Predicting Genome-wide DNA Methylation in Humans

by

Weiwei Zhang

Department of Molecular Genetics and Microbiology
Duke University

Date: _____

Approved:

_____
Barbara Engelhardt, Supervisor

_____
Micah Luftig

_____
Jen-Tsen Ashley Chi

Thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in the Department of Molecular Genetics and Microbiology in the Graduate School of Duke University
2014

Abstract

Predicting Genome-wide DNA Methylation in Humans

by

Weiwei Zhang

Department of Molecular Genetics and Microbiology
Duke University

Date: _____

Approved:

_____
Barbara Engelhardt, Supervisor

_____
Micah Luftig

_____
Jen-Tsen Ashley Chi

An abstract of a thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in the Department of Molecular Genetics and Microbiology
in the Graduate School of Duke University
2014

Abstract

DNA methylation is one of the most studied and important epigenetic modifications in cells, playing a role in DNA transcription, splicing, and imprinting. Recently, advanced genome-wide DNA methylation profiling technologies have been developed, making it possible to conduct methylome-wide association studies. One of the problems with large scale DNA methylation studies is that the current technologies are either targeting only a limited number of CpG sites in the genome or whole genome sequencing is expensive and time consuming for most laboratories. Computational prediction of CpG site-specific methylation levels is the cost-saving and time-saving alternative.

In this work, we found striking patterns of DNA methylation across the genome. We show that correlation among CpG sites decays rapidly within several hundreds base pairs in contrast to the LD structure of genotypes which holds for up to several KB. Using genomic features including, neighbor CpG site methylation and genomic distance, genomic context such as CpG island regions, and genomic regulatory elements, we built random forest classifiers to predict CpG site methylation levels. Our approach achieves 92% prediction accuracy at single CpG sites in different genome-wide methylation datasets. We achieves the highest accuracy as 98% for prediction within CpG island regions. What's more, our method identifies genomic features that interact with DNA methylation, which improves our understanding of mechanisms involved in DNA methylation modification and regulation.

Contents

List of Tables

List of Figures

Acknowledgements

I would like to thank Dr. Barbara Engelhardt and all members of the Engelhardt Lab for their support and guidance. My advisor Dr. Barbara Engelhardt has been supporting me substantially for this thesis work and provided numerous help and guidance. I would also like to thank Dr. Jordana Bell, Dr. Tim Spector, Dr. Panos Deloukas and Dr. Susan Murphy for providing critical data or opinions on this work, which results in a high-quality manuscript that awaits publishing. I would also like to express my gratitute towards Dr. Micah Luftig, Dr. Jen-Tsen Ashley Chi, Dr. Simon Gregory, Dr. Sayan Mukherjee, and Dr. Gregory Crawford for being on my thesis or preliminary exam committee. All your support, together with the support from all of the faculty and staff in the Molecular Genetics and Mocribiogy program, made this thesis work possible.

1

Introduction

Epigenetics is the study of changes in gene expression or complex phenotype that are not associated with changes in DNA sequence. Epigenetics has been shown to be a critical process in cell differentiation, development, and tumorigenesis (Kiefer, 2007; Tost, 2010). DNA methylation is probably the most studied epigenetic modification of DNA, but our understanding of DNA methylation is still in its infancy. In vertebrates, DNA methylation occurs by adding a methyl group to the fifth carbon of the cytosine residue, mainly in the context of 5-CG-3 dinucleotides, or *CpG sites*, mediated by DNA methyl-transferases (DNMTs) (Cedar, 1988; Jaenisch and Bird, 2003). DNA methylation has been shown to play an important functional role in the cell, including involvement in DNA replication and gene transcription, thus playing a crucial role development, aging, and cancer (Barrero et al., 2010; Wolffe, 1999; Rivenbark et al., 2012; Das and Singal, 2004; Scarano et al., 2005; Cedar and Bergman, 2012).

CpG site are underpresented relative to their expected frequency as a result of being a *mutation hotspot*, where the deamination of methylated cytosines often changes CpG sites into TpG sites (Tost, 2010; Lienert et al., 2011). Although CpG site are

mainly methylated across the mammalian genome (Jones, 2012), there are distinct, mostly unmethylated CG-rich regions termed CpG islands (CGIs) that have a G+C content greater than 50% (Tost, 2010; Lienert et al., 2011; Law and Jacobsen, 2010). CGIs account for $1-2\%$ of the genome, and are often located in promoters and exonic regions in mammalian genomes (Shen et al., 2007; Larsen et al., 1992). These CGIs are shown to co-localize with DNA regulatory elements such as transcription factor binding sites (TFBSs) (Brandeis et al., 1994; Macleod et al., 1994; Dickson et al., 2010; Teschendorff et al., 2009; Deaton and Bird, 2011; Choy et al., 2010; Gebhard et al., 2010; Stirzaker et al., 2004) and DNA binding insulator proteins, such as CTCF, which protect downstream DNA from upstream methylation activities (Valenzuela and Kamakaka, 2006). Furthermore, DNA methylation levels have been shown to correlate with gene regions, active chromatin marks (Weber et al., 2007; Meissner et al., 2008; Hawkins et al., 2010), cis-acting DNA regulatory elements, and proximal sequence elements (Shen et al., 2007; Das et al., 2006).

The non-uniform distribution of CpG sites across the human genome and the important role of methylation in cellular processes indicate that genome-wide DNA methylation patterns are needed to fully explore the regulatory mechanism of this critical process (Laird, 2010). Recent advances in methylation-specific microarray and sequencing technologies have enabled the assay of DNA methylation patterns genome-wide and at single base-pair resolution. The current gold standard to asses single-base-pair DNA methylation patterns across the genome within an individual is whole genome bisulfite sequencing (WGBS), which quantifies DNA methylation levels at $\sim$ 26 million (out of 28 million total) CpG sites in the human genome (Laurent et al., 2010; Hon et al., 2011; Lister et al., 2009). However, WGBS is expensive and time-consuming for large samples and can be subject to bisulfite conversion bias. As an alternative, methylation-specific microarrays, and the Illumina HumanMethylation 450K Beadchip in particular, measure bisulphite treated DNA methylation

levels at approximatedly 485,000 preselected CpG sites (Bibikova et al., 2011); however, these arrays assay less than 2% of CpG sites, and this percentage is biased to gene regions. Therefore, quantitative methods are needed to predict methylation status at unassayed sites and genomic regions.

Genome-wide methylation analyses have enabled the study of global methylation patterns of the human genome at single CpG site resolution. In this study, we examined measurements of methylation levels in 100 individuals using the Illumina 450K Beadchip, and one individual using whole genome bisulfite sequencing. Within these methylation profiles, we examined the methylation patterns and correlation structure of these CpG sites, with attention to characterizing methylation patterns in CGI regions. Using features that include neighboring CpG site methylation status and other genomic attributes, we developed a random forest classifier to predict single CpG site methylation status. Using our classifier, we are able to identify DNA regulatory elements that may affect DNA methylation (or vice versa), providing hypotheses for experimental studies on mechanisms by which methylation leads to biological changes or disease phenotypes.

1.1 Related work in DNA methylation prediction

A number of methods to predict methylation status have been developed; see Appendix A for a complete list. Across these methods, a common simplifying assumption is that methylation status is a binary variable, e.g., a CpG site is either methylated or unmethylated in an individual (Bhasin et al., 2005; Bock et al., 2006; Das et al., 2006; Fang et al., 2006; Kim et al., 2008; Fan et al., 2008; Lu, 2010; Zheng et al., 2013). These methods have often limited their predictions to specific regions of the genome, such as CpG Islands (Bock et al., 2006; Kim et al., 2008; Fang et al., 2006; Fan et al., 2008; Previti et al., 2009; Zheng et al., 2013). More broadly, all of these methods make predictons of average methylation status for windows of the genome instead of individual CpG sites.

The majority of these methods are based on a support vector machine (SVM) classifier (Bhasin et al., 2005; Bock et al., 2006; Das et al., 2006; Fang et al., 2006; Fan et al., 2008; Previti et al., 2009; Zhou et al., 2012; Zheng et al., 2013). In Previti *et al.*, their decision tree classifier achieved better performance than an SVM-based classifier. Similarly, Kim *et al.* achieved the best prediction performance using a naive Bayes classifier, and Lu *et al.* used word composition-based encoding method.

Across these methods, features that are used for DNA methylation prediction include, with respect to the genomic region for which methylation is being predicted, DNA composition (proximal DNA sequence patterns), predicted DNA structure (e.g., co-localized introns), repeat elements, TFBSs, evolutionary conservation (e.g., *PhastCons* (Siepel et al., 2005)), number of SNPs, GC content, Alu elements, histone modification marks, and functions of nearby genes. Several studies only used DNA composition features for prediction and achieved prediction accuracies ranging from 75% to 87% (Bhasin et al., 2005; Das et al., 2006; Kim et al., 2008; Lu, 2010; Zhou et al., 2012). In addition to DNA composition features, Fan *et al.* added

histone modification marks. Bock *et al.* used $\sim$ 700 features including DNA composition, DNA structure, repeat elements, TFBSs, evolutionary conservation, and number of SNPs; Zheng *et al.* included $\sim$ 300 features including DNA composition, DNA structure, TFBSs, histone modification marks, and function of nearby genes. They achieved prediction accuracy around 90%. Two other studies did not use any DNA composition features: Fang *et al.* used GC content, Alu elements, and TFBSs, while Previti *et al.* used GC content, repetitive sequences, evolutionary conservation, and DNA structure; their accuracy ranged from 85% to 92%.

The relative success of each of these methods depends on the prediction objectives. Many of the studies, including all the studies that achieved high prediction accuracy ($\geqslant$ 90%) (Bock et al., 2006; Fan et al., 2008; Previti et al., 2009; Zheng et al., 2013), only predicted average methylation status within CGIs or DNA fragments within CGIs. Most of the CpG sites in CGIs are unmethylated across the genome (Jones, 2012; Maunakea et al., 2010) – Manuakea *et al.* found that 16% of all CGIs in cells from human brain were methylated using a WGBS approach – so it is not particularly surprising that classifiers limited to these regions perform well. Studies not limiting prediction to CGI uniformly achieved lower accuracies, ranging from 75% to 86%. Only one of these studies predicted methylation as a continuous variable (Zhou et al., 2012). Their methylation level predictions achieved a maximum correlation coefficient of 0.82 and a corresponding root mean square error of 0.20; however, the study was limited to about 400 DNA fragments instead of a genome-wide analysis.

Our DNA methylation classifier differs from the above methods in that it:

- uses a genome-wide approach,
- is at single CpG site resolution,
- predicts DNA methylation levels instead of status;

- is based on a random forest classifier,
- incorporates a diverse set of genomic features, and
- considers the mechanistic interpretation of the features.

We find that these differences substantially improve the performance of the classifier, and also provide biological insights into how methylation regulates, or is regulated by, specific genomic processes.

2

Materials and methods

2.1 DNA methylation data

2.1.1 Methylation450K data

Illumina HumanMethylation450K array data were obtained for 100 unrelated human participants from the TwinsUK cohort (Moayyeri et al., 2012) All participants provided written informed consent in accordance with local ethics research committees. The 100 individuals were adult unselected volunteers and included 97 female and 3 male individuals (age range 27–78). Whole blood was collected and DNA was extracted using standard protocols.

Illumina HumanMethylation450K array (Illumina 450K) measured DNA methylation values for more than $482,000$ CpG sites per individual at single-nucleotide resolution. Genomic coverage includes 99% of reference sequence genes, with an average of 17 CpG sites per gene region distributed across the promoter, 5'UTR, first exon, gene body, and 3'UTR, and 96% of CpG islands (Rechache et al., 2012; Bibikova et al., 2011).

Methylation values for each CpG site are quantified by the term β, which is the fraction of methylated bead signal over the sum of methylated and unmethylated

bead signal:

$$\beta = \frac{\max(Methy, 0)}{\max(Methy, 0) + \max(Unmethy, 0) + \alpha} \tag{2.1}$$

where *Methy* represents the signal intensity of the methylated probe and *Unmethy* represents the signal intensity of the unmethylated probe. The quantity β ranges from 0 (unmethylated) to 1 (fully methylated).

Data quality control was implemented using R (http://www.r-project.org/) (version 2.15.3). We removed $17,764$ CpG sites for which probes mapped to multiple loci in the human genome reference sequence. CpG sites that are SNPs, that had missing values, or that had detection p-values > 0.01 were excluded. Methylation data from probes mapping to the X and Y chromosomes were excluded. We were left with $394,354$ CpG sites from 100 individuals in downstream analyses. The data were controlled for array number, sample position on the array, age, and sex by fitted linear regression model. The sum of residuals and intercepts of each sites were scaled to the original $[0, 1]$ scale by truncating all sites with values larger than 1 to 1 and all sites with values smaller than 0 to 0. Data quality was assessed to identify sample outliers and batch effectss using Principal Component Analysis (PCA) Gabriel and Odoroff (1990), no obvious outlier was identified.

2.1.2 Whole Genome Bisulfite Sequencing data

We downloaded the WGBS data (BED files) from NCBI Gene Expression Omnibus (GEO) ID GSE46644, sample GSM791827 (Ziller et al., 2013; Hodges et al., 2011). CD19+ B cells were purified from peripheral blood collected from one healthy female donor. Bisulfite sequencing and read mapping processes were described in previous work (Hodges et al., 2011). The methylation levels for each CpG site were quantified by the ratio of the number of methylated and the total reads at each CpG site. Only

CpG sites with greater than 5X coverage were included. Methylation level data from the X and Y chromosomes were excluded. After quality control, there were $10,000,890$ CpG sites in the WGBS data. Because we only used a single sample, we did not control for principal components.

2.2 Correlation and PCA

The statistical analyses were implemented using R and Bioconductor (http://www.bioconductor.org/) (version 2.15.3). Methylation correlations between CpG sites were assessed by the absolute value of Pearson's correlation coefficient and mean square error (MSE):

$$MSE = \frac{\sum_{i=1}^{n}(x_{1i} - x_{2i})^2}{n}, \tag{2.2}$$

where x_{1i} and x_{2i} represent the methylation values of the two CpG sites being compared, n represents the total number of CpG sites being compared.

We performed PCA on methylation values of CpG sites by computing the eigenvalues of the covariance matrix of a subsample of CpG sites using the R function `svd`. Among the $378,677$ CpG sites that have complete feature information, $37,868$ sites (every tenth CpG site) were sampled along the genome across all autosomal chromosomes. Absolute value Pearson's correlation was calculated between each feature and first ten PCs. PCA analysis was performed by plotting the PC biplot (scatterplot of first two PCs).

2.3 Random forest and comparison classifier

The random forest classifier is a widely-used classification method that combines the idea of bagging classification trees and randomizing feature subsets (Breiman, 2001). Compared with traditional decision trees, a random forest randomly generates a set

of independent predictors to extend each tree. To determine the most discriminative feature to add to a tree, the random forest algorithm uses a type of bagging, where a bootstrap sample (a random sample without replacement) is selected from training data to grow each tree, while the remaining training data are used to determine the split that yields the best generalization error (where each split is determined using only a random subset of the features). After each tree has been estimated, they perform prediction by voting on the prediction value, and the final prediction is chosen by majority vote but can be interpreted as the proportion of trees with a '1' prediction (Breiman, 1996, 2001).

Random forest classifiers have been shown to produce low-bias trees and have low correlation among individual trees, creating efficient classifiers, particularly on high-dimensional, noisy data. Moreover, both categorical and continuous prediction features are allowed, and the tree structure explicitly captures interactions among features (Díaz-Uriarte and Alvarez de Andrés, 2006). Due to the random selection of predictors at each tree, the performance of the random forest classifier is highly accurate even when many of the features are not predictive, as long as there are sufficient numbers of trees, avoiding the need for rigorous feature selection (Hua et al., 2005). The random forest classifier also enables feature importance to be quantified, allowing us to functionally interpret the relationship between DNA methylation and specific genomic features.

We used the `randomForest` package in R for the implementation of the RF classifier (Liaw and Wiener, 2002) (version 4.6-7). Most of the parameters were kept as default, but `ntree` was set to $1,000$ to balance efficiency and accuracy in our high-dimensional data. We found the parameter settings for the random forest classifier (including the number of trees) to be robust to different settings, so we did not estimate parameters in our classifier. The Gini index, which calculates the total decrease of node impurity (i.e., the relative entropy of the class proportions before

and after the split) of a feature over all trees, was used to quantify the importance of each feature:

$$I(A) = 1 - \sum_{k=1}^{c} p_k^2, \tag{2.3}$$

where k represents the class and p_k is the proportion of sites belonging to class k in node A.

We used the SVM implementation in the `e1071` package in R (Meyer et al., 2012) with a radial basis function kernel. The parameters of the SVM were optimized using grid search. The penalty constant C ranged from $2^{-1}, 2^1, ..., 2^9$ and the parameter γ in the kernel function ranged from $2^{-9}, 2^{-7}, ..., 2^1$. The parameter combination that had the best performance – $\gamma = 2^{-7}$ and $C = 2^3$ – was used to generate the results used in the comparisons.

For k-NN, we used the `knn` function in R, with the number of neighbors equal to the square root of number of samples in the training set.

For the LR classifier, we used the logistic regression classifier implemented in the R base package with the function `glm` and `family = 'binomial'`. We set the threshold for classification to $\hat{\beta}_{i,j} \geqslant 0.5$.

For the NB classifier, we used the `naiveBayes` function in the R `e1071` package.

2.4 Features for prediction

A comprehensive list of 124 features were used in prediction (Table S2). The *neighbors* features were obtained from data from the Methylation 450K Array; The *position* features, including gene coding region category, location in CGIs, and SNPs, were obtained from the Methylation 450K Array Annotation file; DNA recombination rate data was downloaded from HapMap (phaseII_B37, update date Jan19-2011) (Tanaka, 2005); GC content data were downloaded from the raw data used to

encode the gc5Base track on hg19 (update date Apr242009) from the UCSC Genome Browser (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/gc5Base/) (Meyer et al., 2013), integrated haplotype scores (iHS scores) were downloaded from the HGDP selection browser iHS data of *smoothedAmericas* (update date Feb122009) (Voight et al., 2006), and GERP constraint scores were downloaded from *SidowLab GERP++* tracks on hg19 (http://mendel.stanford.edu/SidowLab/downloads/gerp/) (Davydov et al., 2010); *CREs* features: DNAse I hypersensitive sites data were obtained from the DNase-seq data for the GM12878 cell line produced by Crawford Lab at Duke University (UCSC Accession: wgEncodeEH000534, submitted date Mar202009); 79 specific transcription factor binding sites ChIP-seq data were obtained from the narrow peak files from GM12878 cell line that were available before June 2012 from the ENCODE website; 10 histone modifications were obtained from the peak files from the GM12878 cell line that were available before December 2013 from the ENCODE website; 15 chromatin states were obtained from the Broad ChromHMM data from the GM12878 cell line (UCSC Accession: wgEncodeEH000033 submitted data Mar212011) (Good et al., 2004).

Neighboring CpG site methylation status was encoded as "methylated" when the site has a $\beta > 0.5\%$ and "unmethylated" when $\beta < 0.5\%$. For continuous features, the feature value is the value of that feature at the genomic location of the CpG site; for binary features, the feature status indicates whether the CpG site is within that genomic feature or not. DHS sites were encoded as binary variables indicating a CpG site within a DHS site; TFBSs were included as binary variables indicating the presence of a co-localized ChIP-Seq peak; iHS scores, GERP constraint scores and recombination rates were measured in terms of genomic regions; For GC content, we computed the proportion of G and C within a sequence window of 400 bp, as this feature was shown to be an important predictor in previous study (Fang et al., 2006). Among all 124 features, 122 of them (excluding upstream and downstream

neighboring CpG sites' β values) were used for methylation status predictions, and all excluding upstream and downstream neighboring CpG sites' methylation status τ were used for methylation level predictions. When limiting prediction to specific regions, e.g., CGI, we excluded those features from the data.

2.5 Prediction evaluation

Our methylation predictions were at single CpG site resolution. For regional specific methylation prediction, we grouped the CpG sites into either promoter, gene body, intergenic region classes or CGI, CGI shore & shelf, and non-CGI classes according to the Methylation 450K array annotation file, which was downloaded from the UCSC genome browser (Kent et al., 2002). We only counted CpG sites that have both upstream and downstream neighboring sites in same region. Since the array has non-uniformly distributed probes, the density of probe is highly skewed to promoter and CGI region. To reduce the influence of probe density on the performance of regional specific prediction, we only predicted on CpG sites with neighboring sites within $1kb$ distance. The feature of gene coding region category and CGI status were excluded in different gene coding region prediction and different CGI status prediction, respectively.

The classifier performance was assessed by within-genome, cross-sample and cross-platform validation. In within-genome analyses, ten times we sampled $10,000$ random CpG sites from across the genome into the training set, and we tested on all other held-out sites. The prediction performance for a single classifier was calculated by averaging the prediction performance statistics across each of the 10 trained classifiers. We checked the performance with smaller training set of sizes 100, 1000, 2000, 5000 and $10,000$ sites in the same evaluation setup. In subsequent analyses, we set the size of the training set to $10,000$ randomly chosen CpG sites to balance computational performance and accuracy. For genomic region specific methylation

prediction, the number of sites in each catagory are shown in Tabel 4. The validation was repeated n times where n represents the number of $10,000$ sites training sets in corresponding region. Next, we expand our analyses to the whole genome of each of 100 individual. We considered cross-sample validation to evaluate the consistency of methylation pattern in different individuals. In this case, training was performed from $10,000$ CpG sites in one sample, and the trained classifier was used to predict all of the methylated sites in the remaining 99 samples.

In cross-platform prediction and WGBS prediction, we sampled $10,000$ randomly chosen CpG sites from 450K data or CpG sites categorized as *450K sites* in WGBS data as training sets. We tested on $100,000$ randomly chosen CpG sites that were categorized as *450K site* or *non 450K site* in the WGBS data. The prediction performance for a single classifier was calculated by averaging the prediction performance statistics across each of the 10 trained classifiers.

We quantified the accuracy of the results using the specificity, sensitivity (recall), precision, and accuracy (ACC). Note that *truly significant* CpG sites are those that are methlyated, and *truly null* CpG sites are those that are unmethylated in these data. These values were calculated as follows:

$$SP = \frac{TN}{TN + FP} \tag{2.4}$$

$$SE = \frac{TP}{TP + FN} \tag{2.5}$$

$$Precision = \frac{TP}{TP + FP} \tag{2.6}$$

$$Recall = \frac{TP}{TP + FN} \tag{2.7}$$

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \tag{2.8}$$

for true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) for a particular threshold. We computed Receiver Operating Characteristic (ROC) curves, precision-recall curves, area under the ROC curve (AUC), and area under precision-ecall curve (AUPR); the AUC and AUPR reflect the overall prediction performance considering both type I (FPs) and type II errors (FNs) (Bhasin et al., 2005; Fogarty et al., 2005). We used the `ROCR` package in R (Sing et al., 2005).

To estimate continuous methylation levels ($\hat{beta}$), we used the classifier output of prediction probability from the RF classifier directly as an estimate of a specific $\beta \in [0, 1]$. Prediction accuracy was evaluated using Pearson's correlation coefficient and root mean squared error (RMSE).

$$r_{x,y} = \frac{\sum_{j=1}^{p}(x_j - \bar{x})(y_j - \bar{y})}{(p-1) \cdot \sigma_x \cdot \sigma_y} \tag{2.9}$$

$$RMSE_{x,y} = \sqrt{\frac{\sum_{j=1}^{p}(y_j - x_j)^2}{p}} \tag{2.10}$$

where x_j, y_j are the experimental and predicted values for the j^{th} CpG site, respectively, $\bar{x}$, $\bar{y}$ are the means of the experimental and predicted methylation levels, respectively, and σ_x, σ_y are the empirical standard deviations of the experimental and predicted values, respectively.

3

Results

3.1 Characterizing methylation patterns

First we examined the distribution of DNA methylation values, β, at CpG sites on autosomal chromosomes across all 100 individuals (Figure 3.1A). The majority of CpG sites were either hypermethylated or hypomethylated, with 48.2% of sites having $\beta > 0.7\%$ and 40.4% of sites having $\beta < 0.3$. Using a cutoff of $\beta = 0.5$, across the methylation profiles and individuals, 54.8% of these CpG sites are methylated. We observed distinct patterns of DNA methylation levels in different genomic regions (Figure 3.1B). In the following analysis, we identified co-located CGIs using UCSC genome browser (Kent et al., 2002); CGI shores are regions $0 - 2$ kb away from CGIs in both directions and CGI shelves are regions $2 - 4$ kb away from CGIs in both directions (Bibikova et al., 2011). We found that CpG sites in CGIs were mostly hypomethylated and sites in non-CGIs were mostly hypermethylated, while CpG sites in CGI shore regions had variable methylation levels following a U-shape distribution, and CpG sites in CGI shelf regions were highly hypermethylated. These distinct patterns reflect highly context-specific DNA methylation profiles genome-

wide, leading us to perform a more careful analysis of DNA methylation levels at these loci.

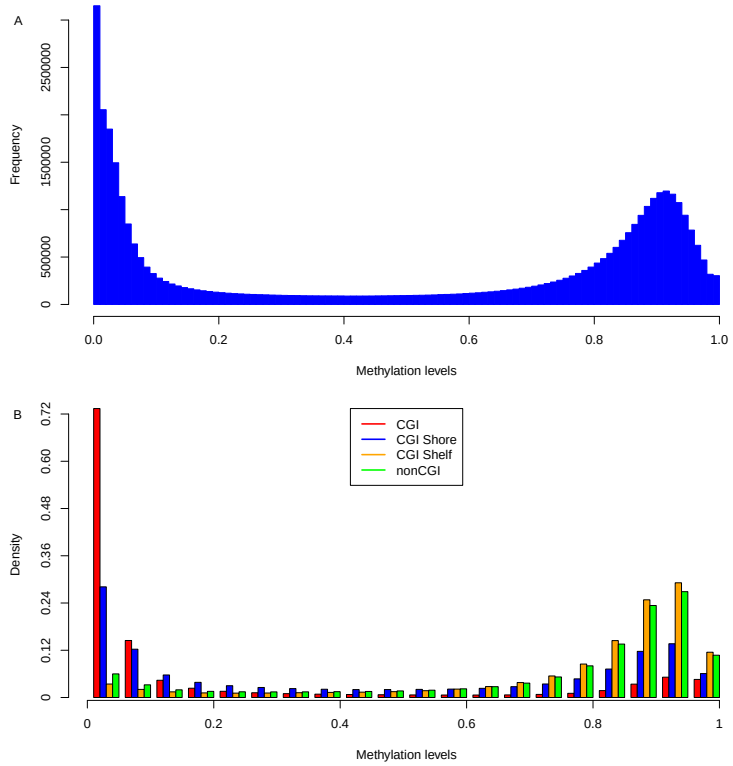

FIGURE 3.1: Distribution of DNA methylation levels at CpG sites across autosomal chromosomes.

Methylation levels across 100 individuals at CpG sites assayed on the 450K array. Panel A: Distribution of DNA methylation values across all CpG sites. Panel B: Distribution of DNA methylation values for CpG sites within CGIs, CGI shores, CGI shelves, and non-CGI regions.

DNA methylation levels at nearby CpG sites have previously been found to be correlated within short genomic regions of $1-2$ kb (Bell et al., 2011; Eckhardt et al., 2006), in contrast with the correlation among genotypes due to linkage disequilibrium (LD) that often extends to large genomic regions from a few kilobases to as much as 100 kb (Reich et al., 2001). We quantified the correlation of methylation levels between neighboring CpG sites in our data using the absolute value of Pearson's correlation. We found that correlation of methylation status between neighboring CpG sites decreased rapidly to approximately 0.4 within about 400 bp, in contrast

to $1-2$ kb in previous studies (Bell et al., 2011; Eckhardt et al., 2006) with sparser CpG coverage (Figure 3.2A). We found the rate of decay in correlation to be highly dependent on genomic context; for example, for neighboring CpG sites in the same CGI shore and shelf region, correlation decreases continuously until it is well below what is expected (Figure 3.2A). Because of the over-representation of CpG sites near CGIs on the array, an increase in correlation can be observed as neighboring sites extend past the CGI shelf regions, where there is lower correlation with CGI methylation levels than we observe in the background.

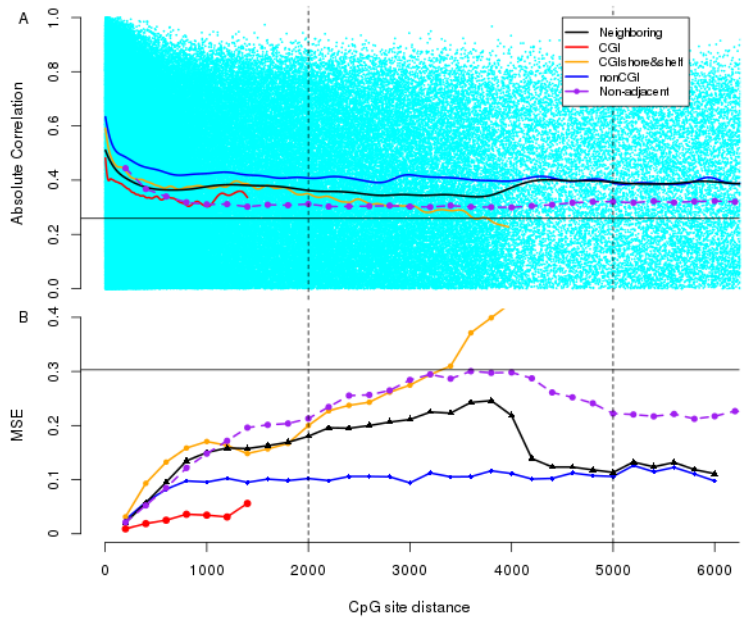

FIGURE 3.2: Correlation of methylation levels between neighboring CpG sites.

The x-axis represents the genomic distance in bases between the neighboring CpG sites, or assayed CpG sites that are adjacent in the genome. Different colors and points represent subsets of the CpG sites genome-wide, including pairs of CpG sites that are not adjacent in the genome but that are the specified distance apart (*non-adjacent*). The CGI shore & shelf CpG sites are truncated at 4000 bp, which is the length of the CGI shore & shelf regions. The solid horizontal line represents the background (absolute value correlation or MSE) levels averaged from $10,000$ pairs of CpG sites from arbitrary chromosomes. Panel A: the absolute value of the correlation between neighboring sites across all individuals (y-axis). The lines represent cubic smoothing splines fitted to the correlation data. Panel B: the MSE was calculated for CpG sites (y-axis) for each pair of CpG sites within the genomic distance window.

We contrast this decay to the *background correlation level*, (0.259), which is the

average absolute value correlation between pairs of randomly selected CpG sites across chromosomes (Figure 3.2A, Figure 3.3). We found substantial differences in correlation between neighboring CpG sites versus arbitrary pairs of CpG sites at identical distances, presumably because of the dense CpG tiling on the 450K array within CGI regions. While correlation decays rapidly within approximately 400 bp between neighboring sites, neighbor correlation does not achieve background correlation levels until several MB, suggesting DNA methylation patterns that expand to large genomic regions.

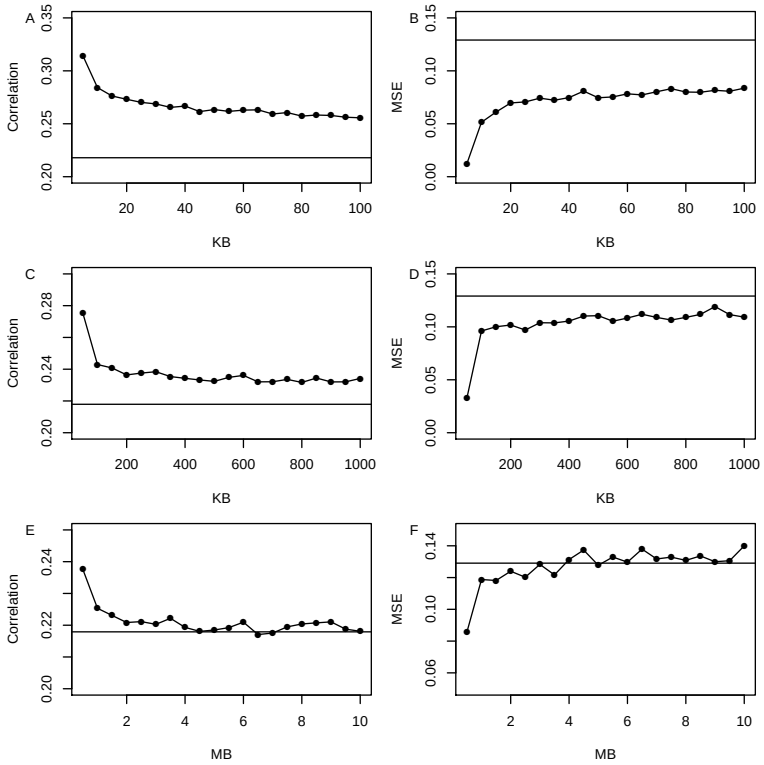

FIGURE 3.3: Correlation and MSE of methylation values between arbitrary pairs of CpG sites.

The x-axes represent the genomic distance between pairs of CpG sites; The left column shows the correlation of CpG sites within 100 kb (Panel A), 1 Mb (Panel C) and 10 Mb (Panel E); the right column plots show the MSE patterns of CpG sites in relation to their genomic distances with distance range 100 kb (Panel B), 1 Mb (Panel D) and 10 Mb (Panel F). The solid horizontal lines represent the background correlation or MSE level calculated from 10000 pairs of CpG sites from a different chromosome.

To quantify the correlation patterns in DNA methylation profiles of neighboring CpG sites within individuals, we calculated the mean square error (MSE) for DNA methylation levels between neighboring CpG sites (Figure 3.2B, Figure 3.3). In general, the MSE trends echo the local patterns seen in the correlation analysis and also appear to be region specific. In CGI regions, the MSE of neighboring sites was low and increased slowly with genomic distance. In contrast, MSE in CGI shore and shelf regions increased rapidly to an MSE higher than background MSE (0.30), indicating that the edges of a single shore and shelf region are less predictive of each other than any two CpG sites at random.

As we observed that methylation patterns at neighboring CpG sites depended heavily on genomic content, we further investigated methylation patterns within CGIs, CGI shores, and CGI shelves. Methylation levels at CGIs and CGI shelves were fairly constant genome-wide and across individuals with CGIs being hypomethylated and CGI shelves being hypermethylated. (Figure 3.4A). In contrast, CpG sites in CGI shores have a monotone increasing pattern of methylation status from CGIs towards CGI shelves, and this pattern is symmetric in the CGI shores upstream and downstream of CGIs. This explains the high levels of variation we observed in methylation values of CpG sites in CGI shore regions If we examine the MSE between pairs of CpG sites' methylation status in these regions, we find that MSE within the CGI and within the CGI shelves is low, consistent with the variance we observed within DNA methylation profiles in these regions (Figure 3.4B). Additionally, we find that the MSE between the CpG sites in the shelves appears to increase as the sites are further away from the CGI on the shelf, suggesting a circular dependency in methylation status between the ends of the shelf sequences.

To quantify the amount of variation in DNA methylation explained by genomic context, we considered the correlation between genomic context and principle components (PCs) of methylation levels across all 100 individuals (Figure 3.5). We found

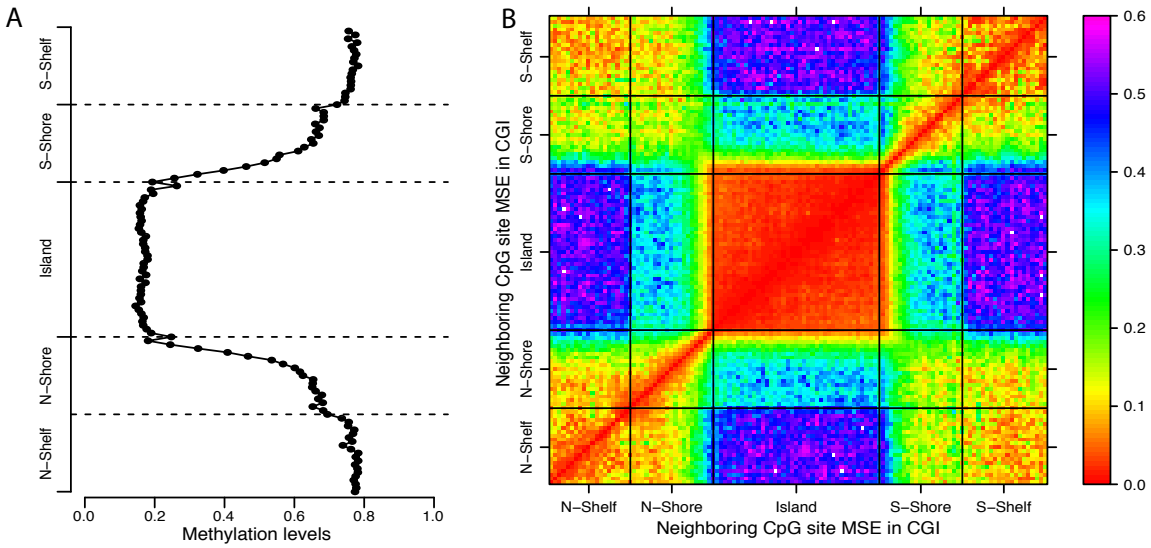

FIGURE 3.4: Methylation structure with respect to CpG islands status.

Since each CGI is a different length, each CGI was split into 40 equal sized windows. Panel A: The points represents the mean β in CGIs, CGI shores or CGI shelves across all sites in all individuals with a window size of 100 bp. Panel B: Methylation levels of each CpG site in each CGI, CGI shore, and CGI shelf were compared with all the other sites in the same CGI region. X-axis and y-axis represent the genomic position of each CGI with a scale of 1:100, i.e. one unit in matrix represents 100 bp distance. The MSE of each unit cell was calculated for all pairwise CpG sites with one site located in the relative scaled position on x-axis and the other one on y-axis, and then averaged over 100 individuals.

that many of the features derived from the CpG site's genomic context appear to be correlated with the first principal component (PC1). Methylation statuses of upstream and downstream neighboring CpG sites and a co-localized DNAse I hypersensitive (DHS) site are the most highly correlated features, both with Pearson's correlation around 0.57 (Figures 3.5). Ten genomic features all have correlation > 0.5 with PC1, including co-localized active TFBSs Elf1 (ETS-related transcription factor 1), MAZ (Myc-associated zinc finger protein), Mxi1 (MAX-interacting protein 1) and Runx3 (Runt-related transcription factor 3), suggesting that they may be useful in predicting DNA methylation status.

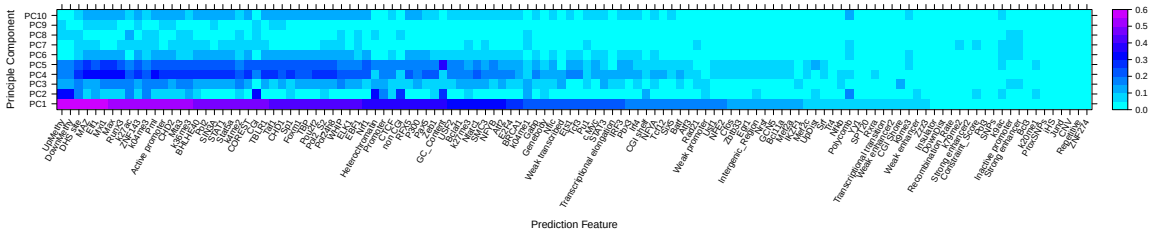

FIGURE 3.5: Correlation matrix of prediction features with first ten principle components of methylation levels.

PCA is performed on methylation levels for $37,865$ CpG sites. The correlation with each of the features was calculated between the first ten methylation PCs (y-axis) and all features (x-axis).

3.2 DNA methylation prediction

3.2.1 Binary methylation status prediction

These observations about patterns of DNA methylation suggest that correlation in DNA methylation is local and dependent on genomic context. Thus, prediction of DNA methylation status based only on methylation levels at neighboring CpG sites may not perform well, especially in sparsely assayed regions of the genome. Using prediction features including neighboring CpG site methylation levels and features characterizing genomic context, we built a classifier to predict DNA methylation status, assuming binary status indicating no methylation (0) or complete methylation (1). For each sample, there were $378,677$ CpG sites with neighboring CpG sites on the same chromosome that we used in these analyses.

The 124 features that we used for DNA methylation status prediction fall into four different classes (see Appendix B). For each CpG site, we include:

- *neighbors*: one upstream and one downstream neighboring (CpG sites assayed on the array and adjacent in the genome) CpG sites' genomic distances, binary methylation status and levels β,

- *genomic position*: binary values indicating co-localization of CpG site with DNA sequence annotations, including promoters, gene body, intergenic region, CGIs, CGI shores and shelves, and nearby SNPs;

- *DNA sequence properties*: continuous values representing the local recombination rate from HapMap (Tanaka, 2005), GC content from ENCODE (Good et al., 2004), integrated haplotype scores (iHS) (Voight et al., 2006), and genomic evolutionary rate profiling (GERP) calls (Davydov et al., 2010).

- *CREs*: binary values indicating CpG site co-localization with cis-regulatory elements (CREs), including DHS sites, 79 specific TFBSs, 10 histone modification marks and 15 chromatin states, all assayed in the GM12878 cell line, the closest match to whole blood (Good et al., 2004)

We built our classifier using a random forest (RF), which is an ensemble classifier that uses a collection of decision trees. We used a cutoff of 0.5 for prediction output, i.e, The prediction with a probability above 0.5 is categrized as methylated site. We quantified prediction accuracy, specificity, sensitivity (recall), precision (1 - false discover rate), area under the Receiver Operating Characteristic (ROC) curve (AUC), and area under the precision recall curve (AUPR) to evaluate our predictions.

Using 122 features (excluding methylaton levels features), we achieved an accuracy of 91.9% and an AUC of 0.96 (Figure 3.6A). We considered the role of each subset of features (Table 3.1). For example, if we only include *genomic position* features, the classifier had an accuracy of 78.6% and AUC of 0.85. Including *DNA sequence properties* and TFBS features increased the accuracy to 85.7% and the AUC to 0.92. When we included all classes of features except for *neighbors*, the classifier achieved an accuracy of 89.0% and an AUC of 0.94, a significant improvement in prediction from only considering *genomic position* features. (t-test; $p = 7.75 \times 10^{-23}$). These results suggest that TFBSs, histone modifications, and chromatin state are

predictive of DNA methylation. However, we also found that the genomic context features improved prediction significantly over using only *neighbors* features, which has an accuracy of 90.7% and an AUC of 0.94 (t-test; $p = 3.45 \times 10^{-18}$).

Table 3.1: Performance of methylation status prediction using different prediction models.
AUC: area under ROC curve; distance: the genomic distance between neighboring CpG sites; gene_pos: genomic position features including gene region status (promoter, gene body, and intergenic region), CGI status (CGI, CGI shore, CGI shelf, and non-CGI), and proximal SNPs; seq_property: DNA sequence properties include GC content, recombination rate, conservation score, integrated haplotype scores; CREs include TFBSs, DHS sites, histone modifications and chromatin state segmentations.

Feature set	Features	Accuracy (%)	AUC	Specificity (%)	Sensitivity (%)
Gene_pos	9	78.6	0.85	72.5	83.6
Gene_pos + seq_property	13	79.5	0.86	71.6	85.9
Gene_pos + seq_property + TFBSs	93	85.7	0.92	78.4	91.7
Gene_pos + seq_property + CREs	118	89	0.94	83.9	93.3
Neighbor + distance	4	90.7	0.94	87.2	93.5
All features	122	91.9	0.96	87.9	95.1

Cross-sample prediction

To determine how predictive methylation profiles were across samples, we quantified the generalization error of our classifier genome-wide across individuals. In particular, we trained our classifier on 10,000 sites from one individual, and predicted methylation status for all CpG sites for the other 99 individuals. The classifier performance was highly consistent across individuals (Figure 3.7).

To test the sensitivity of our classifier to the number of CpG sites in the training set, we investigated the prediction performance for different training set sizes. We found that training sets with greater than $1,000$ CpG sites had fairly similar performance (Figure 3.8). Throughout these experiments, we used a training set size of $10,000$, in order to strike a balance between sufficient numbers of training samples and computational tractability.

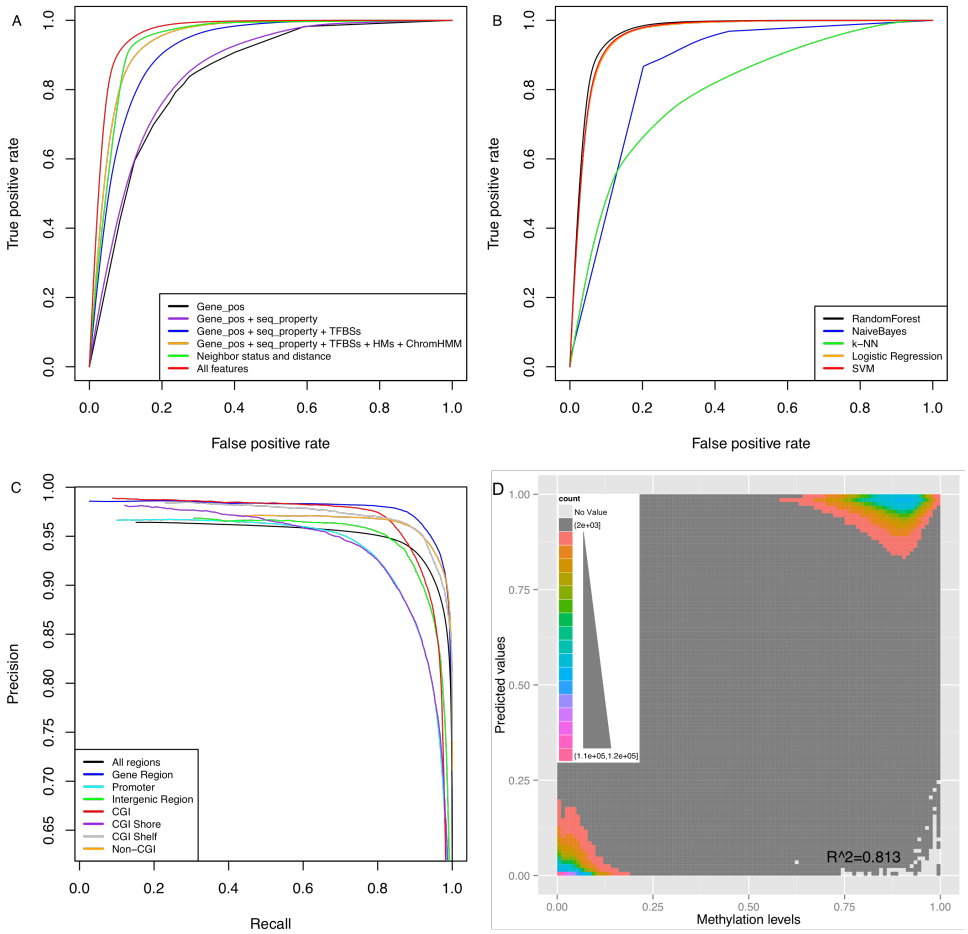

FIGURE 3.6: Prediction performance of methylation status and level prediction.

Panel A, B: ROC curves of within-genome validation of methylation status prediction and prediction with different classifiers. Panel C: Precision-Recall curves of region specific methylation status prediction. Panel D: 2D histogram of predicted methylation levels versus experimental methylation levels distribution.

Cross-platform prediction

To quantify classification across platform and cell type heterogeinity, we investigated the classifier performance on whole genome bisulfite sequencing (WGBS) data (Ziller et al., 2013; Hodges et al., 2011). In particular, we categorized each CpG site in a WGBS sample based on whether that CpG site was assayed on the 450K array (*450K site*) or not (*non 450K site*); *neighboring sites* in the WGBS data are sites that are adjacent on the genome and both *450K sites*. We use one WGBS sample from b-

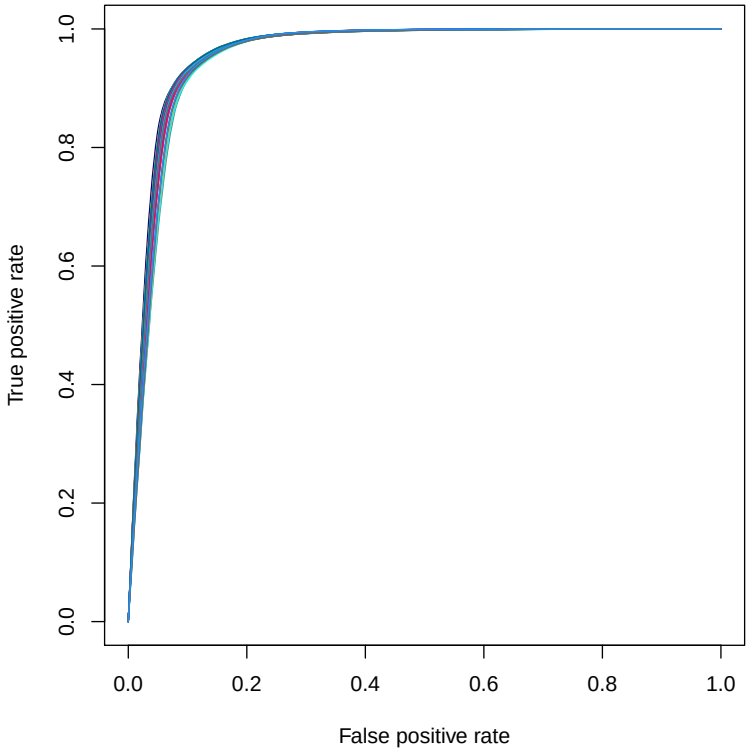

FIGURE 3.7: ROC curves for methylation status prediction.

Each line represents the ROC curve for prediction of CpG sites, training and testing the classifier on data from one individual.

cells, which will match some proportion of each whole blood sample; we note that the 450K array whole blood samples will contain heterogeneous cell types in contrast to the WGBS data. Overall, we see a much higher proportion of hypomethylated CpG sites on the 450K array relative to the WGBS data (Figure 3.9) because of the disproportionate representation of hypermethylated CpG sites within CGIs on the 450K array.

First, we investigated cross-platform prediction, training our classifier on a 450K array sample and testing on WGBS data. We trained the classifier on $10,000$ CpG sites in the 450K array samples, and then we tested on $100,000$ CpG sites twice – once restricting the test set to *450K sites* and once restricting the test set to *non 450K sites* – in WGBS data. We repeated this experiment ten times. Next, we performed the same experiment but trained and tested on the WGBS data. Because

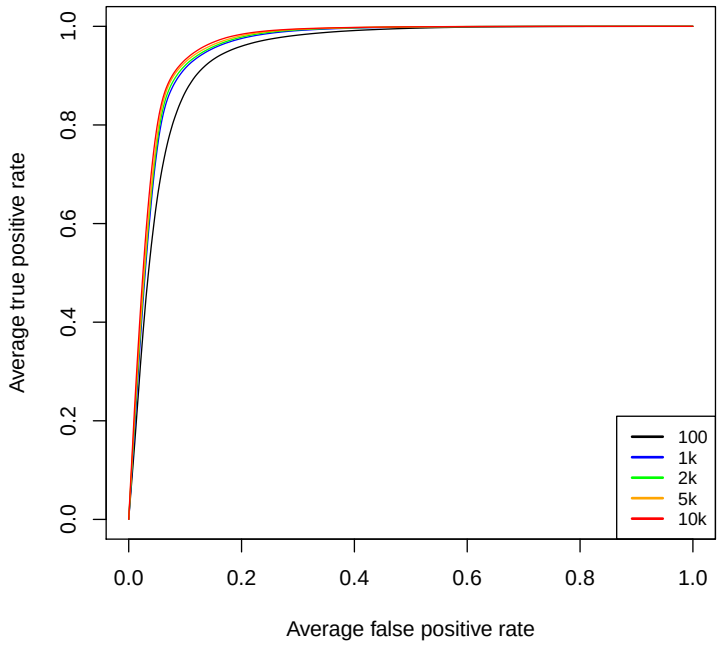

FIGURE 3.8: Generalization error of prediction training size.

Colors represent different training set sizes. For each training set size, the ROC curve is averaged over ten test sets across individuals.

the proportion of hypomethylated and hypermethylated sites was imbalanced for CpG sites not on the 450K array, we used a precision-recall curve instead of a ROC curve to measure the prediction performance (He and Garcia, 2009). We reversed the methylation status in order to assess the quality of the predictions for the less frequent class of hypomethylated CpG sites.

Trained on 450K array data and tested on WGBS *450K sites*, our RF classifier achieved an accuracy of 89.3%; trained on 450K array data and tested on WGBS *non 450K sites*, our RF classifier achieved an accuracy of 92.2% (Figure 3.6; Table 3.2). Training and testing exclusively on WGBS data showed a similar performance, with an accuracy of 90.0% for CpG sites in the *450K sites* and 92.4% for CpG sites in the *non 450K sites* (Figure 3.10). Predictions for CpG sites in *non 450K*

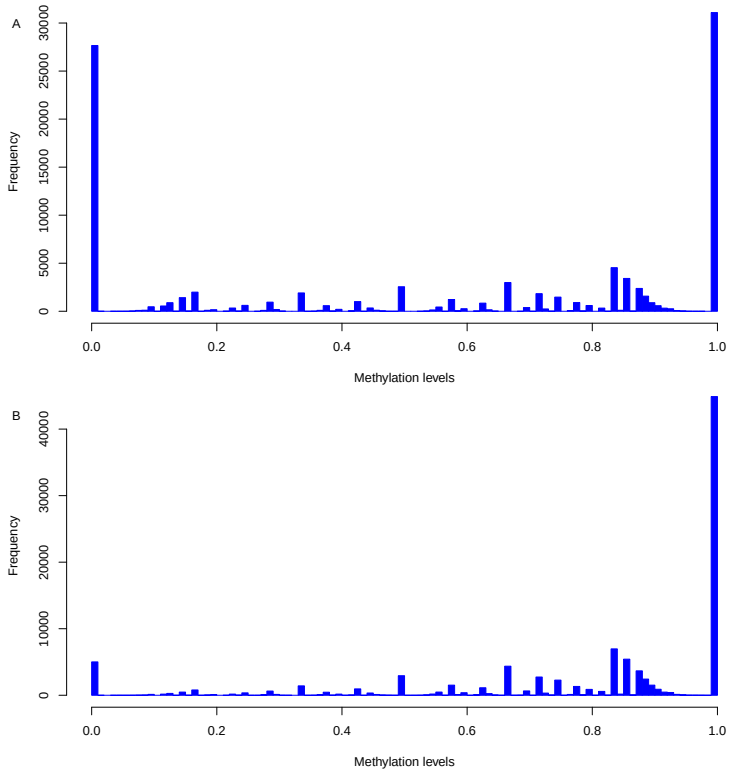

FIGURE 3.9: Distribution of DNA methylation levels at CpG sites from whole-genome bisulfite sequencing data.

Panel A: Distribution of DNA methylation levels across CpG sites from the WGBS data that were categorized as *450K sites*. Panel B: Distribution of DNA methylation levels across CpG sites from the WGBS data that were categorized as *non 450K sites*.

sites had lower precision at high recall rates because it is more difficult to predict unmethylated sites in the sequencing data as there are many more unmethylated CpG sites. These results suggest that our RF classifier is able to generalize across platforms and methylation assay types.

Table 3.2: Performance of methylation prediction using Whole-Genome Bisulfite Sequencing data.

Train set	Test set	Accuracy (%)	Precision (%)	Recall (%)	R	RMSE
WGBS 450K sites	WGBS non 450K sites	92.4	86.5	44.5	0.64	0.24
WGBS 450K sites	WGBS 450K sites	90	91.8	82.3	0.86	0.23
450K data	WGBS non 450K sites	92.2	88.5	41.4	0.62	0.23
450K data	WGBS 450K sites	89.3	93.0	79.3	0.84	0.24

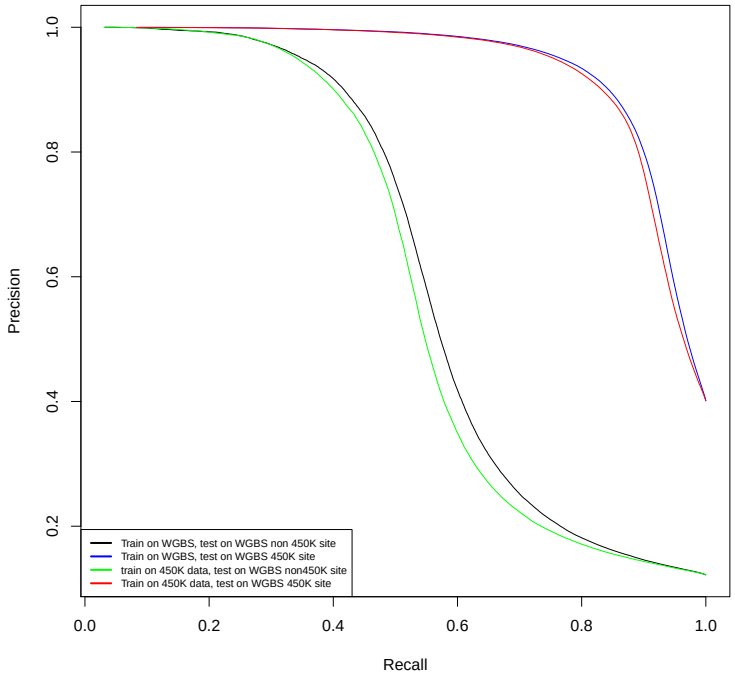

FIGURE 3.10: Prediction performance of Whole Genome Bisulfite Sequencing data.

Presicion-Recall curves were plotted for cross-platform prediction and WGBS prediction. For each category, the curve was generated by averaging the results of all held out test sets.

3.2.2 Comparison of random forest classifier with other classifiers

We compared the prediction performance of our RF classifier with several other classifiers that have been widely used in related work (Table 3.3). In particular, we compared our prediction results from a random forest (RF) classifier with those from a support vector machine (SVM) classifier with a radial basis function (RBF) kernel, a k-nearest neighbors classifier (k-NN), logistic regression (LR), and a naive Bayes classifier (NB). We quantified performance using repeated random resampling, with identical training and test sets across classifiers; we used identical feature sets for all classifiers, including all 122 features used for prediction of methylation status with the RF classifier

We found that k-NN showed the worst performance on this task, with an accuracy

Table 3.3: Performance of methylation status prediction using different classifiers.

Classifier	Accuracy (%)	AUC	Specificity (%)	Sensitivity (%)
k-NN	73.2	0.80	72.6	73.7
Naive Bayes	80.8	0.91	64.4	94.2
Logistic Regression	91.1	0.96	87.3	94.1
SVM	91.3	0.96	86.6	95.1
Random Forest	91.8	0.96	87.9	95.1

of 73.2% and an AUC of 0.80 (Figure 3.6B). The NB classifier showed better accuracy (80.8%) and AUC (0.91). Logistic regression and the SVM classifier both showed good performance, with accuracies of 91.1% and 91.3% and AUCs of 0.96% and 0.96%, respectively. We found that our random forest classifier showed significantly better prediction accuracy than logistic regression (t-test; $p = 3.8 \times 10^{-16}$) and the SVM (t-test; $p = 1.3 \times 10^{-13}$). We note also that the computational time required to train and test the RF classifier was substantially less than the time required for the SVM, k-NN (test only), and NB classifiers. We chose RF classifiers for this task because, in addition to the gains in accuracy over SVMs, we were able to quantify the contribution to prediction of each feature, which we describe below.

3.2.3 Region specific methylation status prediction

Studies of DNA methylation have focused on methylation within promoter regions, restricting predictions to CGIs (Bock et al., 2006; Fang et al., 2006; Fan et al., 2008; Previti et al., 2009; Lu, 2010; Zhou et al., 2012; Zheng et al., 2013); we and others have shown DNA methylation has different patterns in these genomic regions relative to the rest of the genome (Jones, 2012), so the accuracy of these prediction methods outside of these regions is unclear. Here we investigated regional DNA methylation prediction for our genome-wide CpG site prediction method, restricted to CpGs within specific genomic regions (Table 3.4). For this experiment, prediction

was restricted to CpG sites with neighboring sites within 1 kb distance because of the limited size of CGIs.

Within CGI regions, we found that predictions of methylation status using our method had an accuracy of 98.3%. We found that methylation level prediction within CGIs had an $r = 0.94$ and a RMSE of 0.09. As in related work on prediction within CGI regions, we believe the improvement in accuracy is due to the limited variability in methylation patterns in these regions; indeed, 90.3% of CpG sites in CGI regions have a $\beta < 0.5$ (Table 3.4). Conversely, prediction of CpG methylation status within CGI shores had an accuracy of 89.8%. This lower accuracy is consistent with observations of robust and drastic change in methylation status across these regions (Irzarry, 2009; Doi et al., 2009). Prediction performance within various gene regions was fairly consistent, with 94.9% accuracy for predictions of CpG sites within promoter regions, 93.4% accuracy within gene body regions (exons and introns), and 93.1% accuracy within intergenic regions. Because of the imbalance of hypomethylated and hypermethylated sites in each region, we evaluated both the precision-recall curves and ROC curves for these predictions (Figure 3.5C and Figure 3.11).

Table 3.4: Region specific methylation prediction.
Num_sites: Number of CpG sites; Num_sites_1k: Number of sites with both neighbors within 1 kb; Methy%: Percentage of n corresponding regions with neighboring CpG site within 1 kb distance; Num_CV: Number of random subsamples.

Region	Num_sites	Num_sites_1k	Methy%	CV	Accuracy (%)	AUC	Precision (%)	Recall (%)
Promoter	157,468	108,063	0.2070	10	94.98	0.9836	89.06	86.41
Gene body	117,424	35,072	0.6369	3	93.45	0.9741	94.13	95.60
Intergenic	91,177	25,694	0.3960	2	93.05	0.9738	91.23	90.59
CGI	110,612	66,533	0.0973	6	98.32	0.9931	93.94	88.50
CGI shore	89,989	28,232	0.4210	2	89.83	0.9584	88.46	87.15
CGI shelf	36,658	4,736	0.7445	4	89.79	0.9514	89.88	97.54
non-CGI	141,418	34,657	0.7149	3	91.93	0.9489	92.32	96.69
All	378,677	189,735	0.3228	10	91.85	0.9624	90.61	95.06

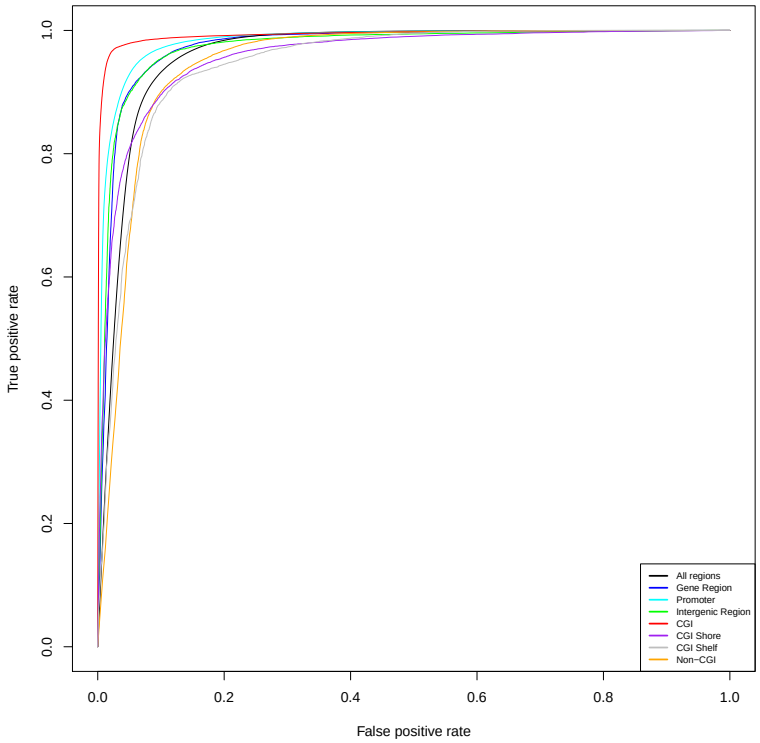

FIGURE 3.11: Prediction performance of region specific methylation status.

ROC curves of region specific methylation status prediction. Colors represent predictions of CpG site methylation status within different genomic regions. For each category, the curve was generated by averaging the results across held out test sets.

3.2.4 Predicting genome-wide methylation levels across platform

CpG methylation levels β in a DNA sample represent the average methylation status across the cells in that sample and will vary continuously between 0 and 1. Since the Illumina 450K array enables precise methylation levels at CpG site resolution in each sample, we used our RF classifier to predict methylation levels at single CpG site resolution. We used the prediction probability as the predicted methylation level. Using all 122 features (excluding methylation status features), we trained our RF classifier on 450K array data and evaluated the Pearson's correlation coefficient (r) and root-mean squared error (RMSE) between experimental and predicted methylation levels (Table 3.5; Figure 3.6D). We found that the experimentally assayed and predicted methylation levels had an $r = 0.90$ and $RMSE = 0.19$. The correlation co-

efficient and the RMSE indicate good recapitulation of experimentally assayed levels using predicted methylation levels across CpG sites.

Table 3.5: Performance of methylation levels prediction using different prediction models.

Feature set	Features	R	RMSE
Gene_pos	9	0.61	0.39
Gene_pos + seq_property	13	0.66	0.34
Gene_pos + seq_property + TFBSs	93	0.80	0.29
Gene_pos + seq_property + CREs	118	0.86	0.23
Neighbor + distance	4	0.87	0.24
All features	122	0.90	0.19

We quantified the performance of methylation level prediction on WGBS data. Trained on CpG sites from the 450K array, and tested the classifier on CpG sites from the WGBS data, both restricted to CpG sites in the *450K sites* set and restricted to CpG sites in the *non 450K sites* set. We achieved different correlation ($r = 0.62, 0.84$, $p < 2.2 \times 10^{-16}$) while much closer RMSE ($RMSE = 0.23, 0.24$, $p = 3 \times 10^{-16}$) when predicting methylation levels for CpG sites in the *450K site* set and CpG sites in the *non 450K site* set, respectively, in WGBS data (Table 3.6).

Table 3.6: Region specific methylation levels prediction.

Region	Num_sites	Num_sites_1k	Methy%	CV	R	RMSE
Promoter	157,468	108,063	0.2070	10	0.9231	0.1346
Gene body	117,424	35,072	0.6369	3	0.9247	0.1681
Intergenic	91,177	25,694	0.3960	2	0.9176	0.1689
CGI	110,612	66,533	0.0973	6	0.9433	0.0926
CGI shore	89,989	28,232	0.4210	2	0.8892	0.1814
CGI shelf	36,658	4,736	0.7445	4	0.8779	0.1871
non-CGI	141,418	34,657	0.7149	3	0.8914	0.1910
All	378,677	189,735	0.3228	10	0.9016	0.1919

3.2.5 Feature importance for methylation prediction

We evaluated the contribution of each feature to overall prediction accuracy, as quantified by the Gini index. The *Gini index* measures the decrease in *node impurity*, or the relative entropy of the observed positive and negative examples before and

after splitting the training samples on a single feature, of a given feature over all trees in the trained RF. We computed the Gini index for each of the 122 features from the trained RF classifier for predicting methylation status. We found that upstream and downstream neighboring CpG site methylation status are the most important features for prediction (Appendix C, Figure 3.12). We found that the feature rankings based on Gini index differed when predicting methylation status in specific genomic regions (Figure 3.7), implying context-specific DNA methylation mechanisms. When we restrict prediction to promoter or CGI regions, the Gini score of the neighboring site status features increased relative to other features, which is consistent with previous results showing high DNA methylation correlation and prediction performance in CGI and promoter resiongs. In contrast, we found that the Gini index of the genomic distance to the neighboring CpG site feature decreased, suggesting that neighboring genomic distance is an important feature to consider for cross-region prediction or prediction cross long genomic distance.

We found that DHS sites are strongly predictive of an unmethylated CpG site; the DHS site feature has the third most significant Gini index across these experiments. This observation is consistent with a previous study showing that CpG sites in DHS sites tend to be unmethylated Tsumagari et al. (2013). CGI status is also an important feature, which is unsurprising given that most CpG sites in CGIs are unmethylated. GC content, which also ranked highly based on Gini index, may have a substantial contribution to prediction as a proxy for other important features, such as CGI status and CpG density. GC content, which was also ranked highly based on Gini index, may have a substantial contribution to prediction as a proxy for other important features, such as CGI status and CpG density.

Several TFs and histone modifications were among the most highly ranked features across experiments, while the specific ranking varies. Some of these CREs have been suggested to be associated with DNA methylation, including Elf1, Runx3,

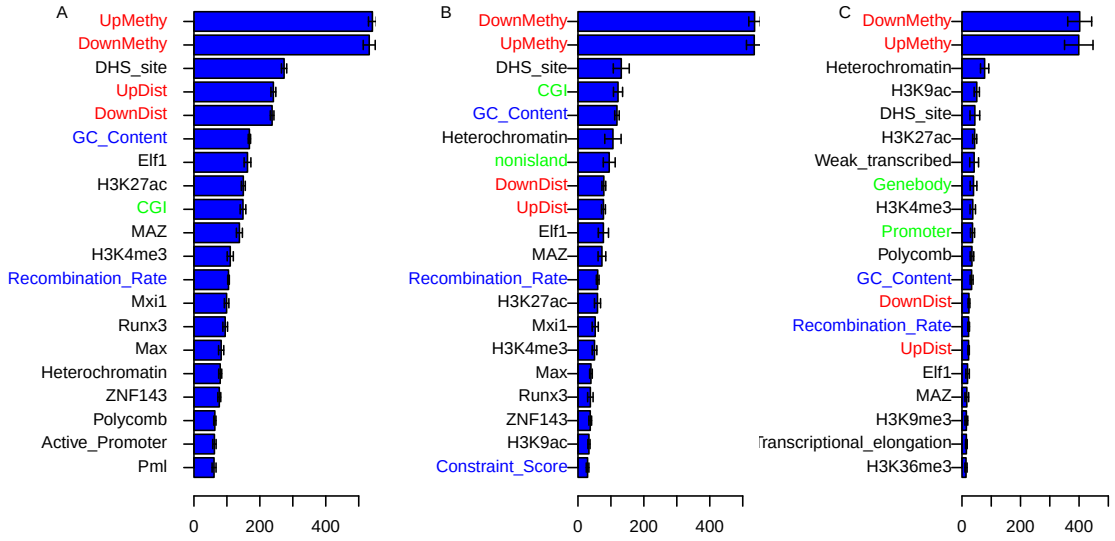

FIGURE 3.12: Top 20 most important features by Gini score.

Gini scores for the top 20 features in predictions in different genomic regions. Panel A: Prediction in any genomic region. Panel B: Prediction in promoter regions. Panel C: Prediction in CGIs. The different colors represent different type of features: *Neighbors* in red, *Gene_Pos* in green, *seq_property* in blue and others in black.

MAZ, Mxi1, and Max. Indeed, the ETS-related transcription factor (Elf1) has been shown to be overrepresented in methylated regions, associating DNA methylation with hematopoiesis in hematopoietic stem cells (Hogart et al., 2012). Runx3 (Runt-related transcription factor 3), a strong tumor suppressor associated with diverse tumor types, has been suggested to be associated with cancer development through regulating global DNA methylation levels (Chuang and Ito, 2010; Li et al., 2002; Kim et al., 2005; Lau et al., 2006; Sato et al., 2006; Weisenberger et al., 2006). Runx3 expression is associated with aberrant DNA methylation in adenocarcinoma cells (Sato et al., 2006), primary bladder tumor cells (Kim et al., 2005), and breast cancer cells (Lau et al., 2006). For another tumor suppressor transcription factor, Mxi1 (MAX-interacting protein 1), expression levels (specifically, lack of expression) have been reported to be associated with promoter methylation levels and neuroblastic tumorigenesis (Lázcoz et al., 2007). It has been suggested that suppression of

MAZ (Myc-associated zinc finger protein) may be associated with DNA methyltransferase I, the key factor for *de novo* DNA methylation (Song et al., 2001, 2003). Mxi1 and MAX (Myc-associated factor X) both interact with c-Myc (myelocytomatosis oncogene), a well characterized oncogene, which has been shown to be methylation sensitive, meaning that the TF motifs contain CpG sites and, thus, TF binding is sensitive to methylation status at those sites (Baron, 2012). This suggests a potential regulatory relationship between MAX, Mxi1, and DNA methylation that may extend to downstream cancer tumor development. The association between specific histone modifications and DNA methylation is poorly understood. A previous study suggested that high H3K4 methylation and Hs acetylation are associated with Myc recognition (Guccione et al., 2006), suggesting regulatory relationships among DNA methylation, histone modification, and transcription factor binding. Our results suggest that further work is needed to investigate the association between DNA methylation and specific histone modification marks.

We found that the correlation between a binary feature and PC1 is proportional to the Gini index of that feature (Figure 3.4 and Appendix D). The variation in the Gini index rankings for CREs varied more than we expected based on the other features (Figure 3.13). CREs that co-occur with CpG sites more often tend to be more important for prediction, according to the Gini index. We found that the Gini index of a binary feature has a log linear relationship with the number of co-occurrences of that binary feature with CpG sites in the data set: the more often a CpG site in the training data co-occurred with a CRE, the higher the Gini index rank of that CpG site (Figure 3.13). There were several outliers to this trend, including status of promoter Pol3 (RNA polymerase III), C-fos (a proto-oncogene), and histone modifications H3K9ac and H4K20me. These features were less important than we would predict using the fitted linear regression model of log Gini index. This trend limits the strong conclusions that associate specific CREs with DNA methylation

biochemically from a high Gini index rank for that CRE; it may be that there are general relationships between CRE and CpG sites that we are learning, but a relatively high CRE frequency in these data will artificially inflate the rank of that CRE in comparison to the others (Figure 3.13). Most CpG sites within TFBSs have low average methylation levels (Appendix D). Several TFBSs have disproportionately high average methylation levels: for example ZNF274 (Zinc-finger protein 274) and JunD (Jun D proto-oncogene); however, both of these outliers also have a low co-occurrence frequency with CpG sites in these data, suggesting that this finding may be an artifact.

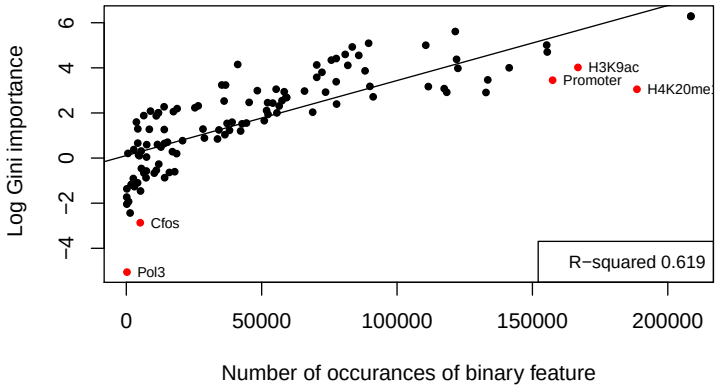

FIGURE 3.13: Correlation of Gini index and the co-occurrence counts of binary features.

The x-axis represents the number of CpG sites co-localized with the corresponding feature (the number of CpG sites that were encoded as '1' for corresponding feature). The y-axis represents the log value of Gini scores of prediction with neighboring CpG site of arbitrary distance. The line was fitted by linear regression. The outliers are highlighted in red. UpMethy: upstream CpG site's methylation status; DownMethy: downstream CpG site's methylation status; Pol3, Cfos, Rxra: TFBSs of Pol3, c-fos and Rxr-α.

4

Discussion

Genome-wide association studies (GWAS) have successfully identified many gene locus associated with common diseases. However, a large percentage of causality still cannot be explained by genetic variation but has been shown to associate with epigenetic factors. As one of the most important epigenetic modification, DNA methylation has been shown to have downstream effects on cell differentiation, embryonic development, aging, and cancer. Although genome-wide DNA methylation profiling technologies are available, most of the genome-wide DNA methylation studies are using a variety of different technologies with limited and varied genomic coverages, which 1) do not characterize the whole genome methylome, 2) introduce difficulties in comparison of different studies and 3) create significant barriers of meta-analysis. Thus, computational prediction of DNA methylation is of great interest for cost and time saving, missing data filling and meta-analysis that facilitate information sharing and collaboration in scientific community.

In this work, we first characterized genome-wide context-dependent DNA methylation pattern. We found that the correlation of DNA methylation decreases rapidly within short genomic distance, making it difficult to predict DNA methylation lev-

els solely using neighboring DNA methylation information. On the other hand, we found that DNA methylation patterns and correlations are highly genomic context-dependent. It has recently been found that CGI shore regions have distinct methylation pattern from CGIs and the high variation in CGI shore regions may relate to tumorigenesis (Irzarry, 2009; Doi et al., 2009). We indeed observed distinct methylation patterns in CGI, CGI shore and CGI shelf regions. In a further step, we found that not only the DNA methylation levels are distinct, but also the correlation between neighboring CpG sites are distinct. The correlation in CGI regions is constantly higher than other regions; the correlation in CGI shore and shelf regions decreases rapidly in relation to genomic distance. Besides, the correlation in CGI shore and shelf regions are also correlated with the relative location of the CpG sites within the CGI shore and shelf, implying that the correlation structure could be useful in modeling DNA methylation and predicting DNA methylation levels.

Based on the DNA methylation patterns and previous findings, we built prediction classifiers to predict DNA methylation levels in humans. We incorporated multiple information as prediction features, including 1) neighboring CpG site methylation levels and genomic distance, 2) genomic position such as CGI regions and gene coding regions, 3) DNA sequence property such as recombination rate and GC content, and 4) regulatory elements such as transcription factor binding sites and histone modification marks. By using 122 genomic features, we could predict both DNA methylation status and levels with high accuracy cross genome and samples using both genome-wide microarray and bisulfite sequencing data. Our high prediction performance of cross platform prediction implied that our classifiers is highly applicable to enable meta-analysis that consolidate results from different methylation techniques and platforms. It could be applied to expand previous studies using techniques with low measurement density.

Except our high accuracy, our classifiers allow us to quantify the contribution

of each feature to DNA methylation prediction. Except neighboring site methylation levels, we identified several features that are highly predictive, including co-localization of DNase I Hypersensitive sites, GC content, and several TFBSs and histone modification marks such as Elf1, H3K27ac, H3K4me3, Mxi1 and Max. Our classifier could not only predict DNA methylation but also suggest genomic features that may regulate, or be regulated by, DNA methylation. These results imply important regulatory mechanisms of DNA methylation in human cells.

Appendix A

Appendix A

The table shows related work on DNA methylation prediction. MethDB is a database of measurements of DNA methylation from variety of studies and methods (Amoreira et al., 2003), which is regularly updated. HEP: Human Genome Project, which contains about 1.9 million CpG methylation values from chromosomes 6, 20 and 22 from 12 cell types across 43 samples.

	Data set	Classifier	Features	Training set	Prediction	Best or overall performance
Bhasin *et al.* 2005	MethDB with methylation value of 5000 DNA fragments	SVM	DNA composition	Methylation status of DNA fragment with window size of 9-89bp	Methylation Status	ACC: 75%, AUC: 0.82, MCC 0.504
Bock *et al.* 2006	Measurements of CGI methylation on human Chromosome 21 (Yamada et al. 2004) with methylation status of 149 CGIs	SVM	DNA composition, predicted DNA structure, repeat element, TFBSs, evolutionary conservation elements, number of SNPs	CGIs methylation status	Methylation status	ACC: 91.5%
Das *et al.* 2006	Human brain data (Rollin et al. 2006) with methylation status of 5000 DNA domains	SVM (K-means clustering, linear discriminant analysis, logistic regression)	DNA composition	Methylation status of DNA fragments with window size of 800bp	Methylation status	ACC: Overall: 86%; CGIs: 96.5%; nonCGIs: 84%
Fang *et al.* 2006	Human brain data (Rollin et al. 2006) with methylation status of 4000 DNA domains	SVM	GC content, Alu element, TFBSs	Methylation status of CG-rich regions with window size of 100-400 bp	Methylation status	ACC: 84.52% CC: 0.686
Kim *et al.* 2007	Sequencing data on 25 gene-related CGIs	Naive Bayes (SMO, Multilayer Perception, Instance Based Classifier)	DNA composition	CGI methylation status of DNA fragments with window size of 30bp	Methylation status	ACC: over 75%
Fan *et al.* 2008	HEP (Eckhardt et al. 2006) with methylated status of 500 DNA fragments of T cell	SVM	DNA composition, histone methylation marks	CGIs methylation status	Methylation status	ACC: 89.94%
Previti *et al.* 2009	HEP with methylated status of 400-500 DNA fragments in 12 cell types	Decision tree (SVM)	GC content, repetitive sequences, evolutionary conservation, DNA structure	CGI methylation status	Methylation status	CC: 0.775 ACC: 91.67%
Lu *et al.* 2010	HEP with methylated status of 3500 DNA fragments of CD4 lymphocytes	Word composition based encoding method	DNA composition	Methylation status of DNA fragments with window size of 1000 bp	Methylation status	ACC: 77.45%

	Data set	Classifier	Features	Training set	Prediction	Best or overall performance
Zhou *et al.* 2012	MethDB with methylation value of 400 human gene fragments	SVM	DNA composition	Methylation level of DNA fragments with window size of 19-129 bp	Methylation status and levels	Methylation status prediction: ACC: 0.82, MCC: 0.64; methylation level prediction: R: 0.82, RMSE: 0.2042
Zheng *et al.* 2013	HEP with methylated status of 400-500 DNA fragments from 12 cell types	SVM	DNA composition, Conserved TFBSs, DNA structure, functional role of nearby genes and histone marks	CGIs methylation status	Methylation status	ACC: ranges from 72% to 95% with different feature combinations in different cell types
Ma *et al.* 2014	Predicting DNA methylation level across human tissues	SVM with RBF kernel	Methylation levels in one cell type at same CpG site	Methylation levels of CpG sites in different tissues	Methylation levels	Methylation level prediction: r^2: 0.89-0.95

Appendix B

Appendix B

List of features prediction feature. Columns include the category of features, the source of the data, and the name of the features.

Category	Source	Feature Name
Neighbors	Methylation 450K data	Upstream neighboring CpG site methylation status & level
		Downstream neighboring CpG site methylation status & level
	Distance features	Upstream neighboring distance
		Downstream neighboring distance
Genomic Position	Methylation 450K annotation	SNPs present within probe > 10bp from query site
		SNPs present within probe < 10bp from query site
	UCSC database knownGene.txt.gz	Presence in promoter
		Presence in gene body
		Presence in intergenic region
	UCSC database cpgIslandExt.txt.gz	Presence in CpG island
		Presence in CGI shore
		Presence in CGI shelf
		Presence in non-CGI region
DNA seq property	Genome Evolutionary Rate Profiling (GERP)	Constraint Score
	hgdp selection brower smoothedAmericas.iHS.gff	Integrated Haplotype Score (iHS)
	HapMap recombination/2011-01_phaseII_B37/	Recombination Rate
	UCSC goldenPath hg19.gc5Base.txt.gz	GC Content

Cis-regulatory elements	ENCODE UCSC Accession wgEncodeEH000534	DNAse I hypersensitive sites
	ENCODE Track Name TFBS, narrowPeak, Cell type: GM12878, Date freeze before July 2012	CTCF
		Cfos
		E2F4
		EBF1
		ELK1
		GCN5
		IKZF1
		IRF3
		Jund
		MAZ
		Max
		Mxi1
		NFE2
		NFYA
		NFYB
		Nfkb
		Nrf1
		P300
		Pol2
		Pol2_S2
		Pol3
		RFX5
		Rad21
		SIN3A
		SMC3
		SPT20
		STAT1
		STAT3
		TBLR1
		TBP
		Tr4
		USF2
		WHIP
		Yy1
		ZNF143
		ZNF274
		Zzz3
		Atf2
		Atf3
		Batf
		Bclaf1
		Bcl11a
		Bcl3
		Egr1
		Elf1
		Ets1
		Foxm1

	Gabp
	Irf4
	Mef2a
	Mef2c
	Mta3
	Nfatc1
	Nfic
	Nrsf
	Pax5
	Pbx3
	Pml
	Pol24h8
	Pou2f2
	Pu1
	RFX5
	Runx3
	Rxra
	Six5
	Sp1
	Srf
	Stat5a
	Taf1
	Tcf12
	Tcf3
	Usf1
	Zbtb33
	Zeb1
	c.Myc
ENCODE Track name histone modification, Peak, Cell type: GM12878	H3K4me1
	H3K04me3
	H3K4me2
	H3K9ac
	H3K9me3
	H3K27ac
	H3K27me3
	H3K36me3
	H3K79me2
	H4k20me1
UCSC Accession wgEncodeEH000033 wgEncodeBroadHmmGm12878HMM.bed	Active promoter
	Weak promoter
	Inactive promoter
	Strong enhancer
	Strong enhancer2
	Weak enhancer
	Weak enhancer2
	Insulator
	Transcriptional transition
	Transcriptional elongation
	Weak transcribed

	Polycomb
	Heterochromatin
	Repetitive
	CNV

Appendix C

Appendix C

Gini importance scores for all features. Gini_all: Gini scores for prediction in any genomic regions; Gini_promoter: Gini scores for prediction in promoter regions; Gini_CGI: Gini scores for prediction in CGIs.

Feature Name	Gini_all	Gini_promoter	Gini_CGI
Upstream neighbor CpG site methylation status	541.552	534.886	399.055
Downstream neighbor CpG site methylation status	531.864	535.967	402.254
DHS site	273.547	131.399	44.035
Upstream neighbor CpG site distance	241.021	77.379	22.678
Downstream neighbor CpG site distance	237.204	78.518	23.936
GC content	168.279	118.312	32.919
Elf1	162.508	77.157	19.12
H3K27ac	149.77	59.323	43.451
Within CGI	148.856	121.591	NA
MAZ	137.958	72.734	16.621
H3K4me3	110.184	50.229	36.9
Recombination rate	104.877	59.93	22.709
Mxi1	99.106	52.39	8.424
Runx3	94.866	37.644	7.971
Max	82.612	39.422	9.318
Heterochromatin	79.684	106.136	77.419
ZNF143	76.93	37.272	8.703
Polycomb	63.44	27.453	34.14
Active promoter	62.11	24.761	5.7
Pml	61.002	27.392	5.936
H3K9ac	55.64	33.347	50.589

Presence of non-CGI	54.75	94.865	NA
H3K36me3	53.501	22.92	13.819
iHS	47.941	28.896	10.067
Mta3	47.802	18.679	4.93
CHD2	44.864	27.873	5.476
Constraint score	42.809	29.204	13.145
BHLHE40	35.904	20.189	4.537
H3K9me3	32.004	18.461	15.002
Presence of promoter	31.598	NA	36.049
Pol2	29.515	17.188	3.097
Weak transcribed	25.533	25.837	41.617
Presence of CGI shelf	25.461	6.503	NA
Presence of CGI shore	24.028	20.577	NA
H3K4me2	23.788	12.885	6.843
Presence of gene	21.74	NA	39.436
SIN3A	21.18	12.887	2.066
H4K20me1	21.122	13.009	5.988
CTCF	19.86	8.4	2.733
Stat5a	19.513	8.783	2.268
STAT1	18.935	10.579	1.749
H3K79me2	18.573	11.452	5.647
H3K27me3	18.487	12.541	4.745
H3K04me1	18.311	11.599	5.42
Presence in intergenic region	15.118	NA	4.743
TBLR1	14.654	7.967	1.551
COREST	14.605	8.455	1.497
WHIP	12.743	6.98	0.507
SMC3	12.541	4.34	1.252
SNPs present within probe $>$ 10bp from query site	11.814	6.986	2.629
Taf1	11.667	8.857	1.414
CHD1	11.445	6.645	1.178
Pol2_S2	10.963	10.742	2.869
SNPs present within probe $<$ 10bp from query site	10.129	5.429	2.292
EBF1	10.062	6.626	1.02
Weak enhancer	9.716	6.95	3.011
Weak promoter	9.302	6.965	2.074
Transcriptional elongation	8.955	3.349	14.994
Sp1	8.231	8.698	2.218
Weak enhancer2	8.018	6.15	1.685
Pu1	7.836	2.423	0.327
Pol24h8	7.65	6.756	1.829
Inactive promoter	7.448	6.177	1.147
Pou2f2	7.413	6.137	1.42
Strong enhancer	7.401	6.207	1.184
Foxm1	6.941	5.752	0.924

Insulator	6.57	2.888	2.127
Rad21	6.496	1.607	0.399
TBP	5.239	5.528	0.605
Strong enhancer2	4.939	2.177	0.486
P300	4.913	3.742	0.424
ELK1	4.727	3.497	0.434
Zeb1	4.655	4.064	1.558
Pax5	4.565	2.944	1.251
Transcriptional transition	3.65	1.876	1.74
NFYB	3.617	3.331	0.643
IKZF1	3.583	1.85	0.07
Batf	3.542	1.382	0.091
USF2	3.48	2.833	0.502
Nfatc1	3.422	3.083	0.725
Nrf1	3.32	3.734	0.905
RFX5	2.818	2.333	0.321
Atf2	2.431	1.831	0.184
Bclaf1	2.333	2.099	0.366
Nfic	2.162	1.299	0.076
STAT3	2.023	1.503	0.053
Nfkb	1.94	1.316	0.036
Irf4	1.91	1.25	0.188
Tcf12	1.827	0.999	0.355
Bcl11a	1.813	0.951	0.224
Pbx3	1.631	0.801	0.123
Bcl3	1.458	0.569	0.651
Mef2a	1.365	0.725	0.059
Tcf3	1.328	0.679	0.409
Repetitive	1.23	0.743	0.426
BRCA1	1.218	0.829	0.213
Nrsf	1.147	0.783	0.288
Mef2c	1.109	0.663	0.027
Usf1	1.047	0.775	0.314
c.Myc	0.764	0.697	0.299
Zbtb33	0.631	0.402	0.095
IRF3	0.583	0.592	0.024
Six5	0.564	0.778	0.029
E2F4	0.546	0.595	0.098
Gabp	0.529	0.616	0.208
NFE2	0.515	0.426	0.042
NFYA	0.512	0.404	0.045
Ets1	0.418	0.635	0.116
Atf3	0.418	0.525	0.034
Tr4	0.404	0.431	0.005
GCN5	0.334	0.314	0.045

Yy1	0.31	0.432	0.026
SPT20	0.307	0.326	0.044
Srf	0.282	0.181	0.103
CNV	0.256	0.106	0.21
Egr1	0.232	0.35	0.122
Jund	0.178	0.177	0
Zzz3	0.145	0.246	0.017
ZNF274	0.13	0.024	0
Rxra	0.088	0.012	0.027
Cfos	0.057	0.091	0.011
Pol3	0.006	0.002	0

Appendix D

Appendix D

Number of occurrences and methylation levels of binary features. Columns are Feature name, Feature counts (or the number of CpG sites that co-occur with these features in our data), and percentage of methylation CpG sites (or the proportion of CpG sites that co-occur with these features that are methylated).

Feature Name	Feature counts	Percentage of methylated CpG sites
Upstream neighboring CpG site methylation status	208536	0.841
Downstream neighboring CpG site methylation status	208535	0.841
SNPs present within probe > 10bp from query site	45374	0.578
SNPs present within probe < 10bp from query site	26557	0.64
DHS site	121516	0.111
Presence in CGI	110612	0.161
Presence in CGI Shore	89989	0.468
Presence in CGI Shelf	36658	0.864
Presence in non-CGI	141418	0.819
Presence in Promoter	157468	0.314
Presence in Gene body	117424	0.729
Presence in Intergenic region	91177	0.673
Atf2	28854	0.067
Atf3	7280	0.017
BHLHE40	70342	0.05
BRCA1	18594	0.023
Batf	14022	0.162
Bcl11a	7539	0.163
Bcl3	2759	0.356
Bclaf1	33688	0.05
CHD1	54108	0.041
CHD2	72260	0.042
COREST	58915	0.042
CTCF	48396	0.071
Cfos	5154	0.004
E2F4	17883	0.011
EBF1	56467	0.089
ELK1	44385	0.021
Egr1	5257	0.016

Elf1	89501	0.048
Ets1	14155	0.01
Foxm1	52346	0.043
GCN5	4176	0.023
Gabp	15904	0.007
IKZF1	8503	0.208
IRF3	11087	0.023
Irf4	14074	0.098
Jund	155	0.468
MAZ	83459	0.035
Max	77542	0.037
Mef2a	5499	0.12
Mef2c	4813	0.125
Mta3	88190	0.101
Mxi1	80851	0.042
NFE2	6491	0.039
NFYA	10329	0.018
NFYB	28315	0.042
Nfatc1	38120	0.089
Nfic	20773	0.084
Nfkb	4235	0.183
Nrf1	42227	0.018
Nrsf	4514	0.023
P300	39077	0.043
Pax5	42760	0.051
Pbx3	12800	0.044
Pml	81813	0.07
Pol2	77520	0.086
Pol24h8	68843	0.124
Pol2_S2	77669	0.14
Pol3	255	0.006
Pou2f2	55632	0.055
Pu1	17400	0.082
RFX5	36423	0.025
Rad21	11054	0.1
Runx3	85861	0.067
Rxra	1410	0.029
SIN3A	55349	0.014
SMC3	36131	0.08
SPT20	1867	0.112
STAT1	58324	0.03
STAT3	15220	0.083
Six5	7402	0.025
Sp1	51817	0.03
Srf	2987	0.043
Stat5a	65821	0.067
TBLR1	59257	0.045
TBP	50953	0.041
Taf1	52199	0.015
Tcf12	11512	0.083
Tcf3	17033	0.051
Tr4	2589	0.061
USF2	34220	0.048
Usf1	7466	0.034
WHIP	57505	0.092
Yy1	1827	0.044
ZNF143	75729	0.043
ZNF274	194	0.696
Zbtb33	5558	0.048
Zeb1	37222	0.033
Zzz3	838	0.049
c.Myc	11989	0.019
H3K4me1	132850	0.369
H3K4me3	155476	0.235
H3K4me2	111551	0.210

H3K9ac	166816	0.52
H3K9me3	133511	0.51
H3K27ac	155274	0.225
H3K27me3	118352	0.325
H3K36me3	122475	0.191
H3K79me2	73603	0.507
H4K20me1	188613	0.524
Active promoter	70341	0.041
Weak promoter	25332	0.236
Inactive promoter	11843	0.43
Strong enhancer	11843	0.43
Strong enhancer2	3720	0.714
Weak enhancer	13960	0.375
Weak enhancer2	8930	0.784
Insulator	6438	0.367
Transcriptional transition	4255	0.876
Transcriptional elongation	18777	0.956
Weak transcribed	35266	0.918
Polycomb	41145	0.36
Heterochromatin	121981	0.881
Repetitive	614	0.480
CNV	210	0.567

Bibliography

Amoreira, C., Hindermann, W., and Grunau, C. (2003), "An improved version of the DNA methylation database (MethDB)," *Nucleic Acids Research*, 31, 75–77.

Baron, B. (2012), "Chapter 1. Breaking the Silence : The Interplay Between Transcription Factors and DNA Methylation," in *Methylation - From DNA, RNA and Histones to Disseases and Treatment*, pp. 3–26, InTech.

Barrero, M. J., Boué, S., and Izpisúa Belmonte, J. C. (2010), "Epigenetic mechanisms that regulate cell identity." *Cell stem cell*, 7, 565–70.

Bell, J. T., Pai, A. a., Pickrell, J. K., Gaffney, D. J., Pique-Regi, R., Degner, J. F., Gilad, Y., and Pritchard, J. K. (2011), "DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines." *Genome biology*, 12, R10.

Bhasin, M., Zhang, H., Reinherz, E. L., and Reche, P. a. (2005), "Prediction of methylated CpGs in DNA sequences using a support vector machine." *FEBS letters*, 579, 4302–8.

Bibikova, M., Barnes, B., Tsan, C., Ho, V., Klotzle, B., Le, J. M., Delano, D., Zhang, L., Schroth, G. P., Gunderson, K. L., Fan, J.-B., and Shen, R. (2011), "High density DNA methylation array with single CpG site resolution." *Genomics*, 98, 288–95.

Bock, C., Paulsen, M., Tierling, S., Mikeska, T., Lengauer, T., and Walter, J. (2006), "CpG island methylation in human lymphocytes is highly correlated with DNA sequence, repeats, and predicted DNA structure." *PLoS genetics*, 2, e26.

Brandeis, M., Frank, D., Keshet, I., Siegfried, Z., Mendelsohn, M., Nemes, A., Temper, V., Razin, A., and Cedar, H. (1994), "Sp1 elements protect a CpG island from de novo methylation." *Nature*, 371, 435–8.

Breiman, L. (1996), "Bagging predictors," *Machine Learning*, 24, 123–140.

Breiman, L. (2001), "Random forests," *Machine learning*, pp. 5–32.

Cedar, H. (1988), "DNA methylation and gene activity." *Cell*, 1964, 93–124.

Cedar, H. and Bergman, Y. (2012), "Programming of DNA Methylation Patterns," *Annual Review of Biochemistry*, 81, 97–117.

Choy, M.-K., Movassagh, M., Goh, H.-G., Bennett, M. R., Down, T. a., and Foo, R. S. Y. (2010), "Genome-wide conserved consensus transcription factor binding motifs are hyper-methylated." *BMC genomics*, 11, 519.

Chuang, L. S. H. and Ito, Y. (2010), "RUNX3 is multifunctional in carcinogenesis of multiple solid tumors." *Oncogene*, 29, 2605–2615.

Das, P. M. and Singal, R. (2004), "DNA methylation and cancer." *Journal of Clinical Oncology*, 22, 4632–42.

Das, R., Dimitrova, N., Xuan, Z., Rollins, R. a., Haghighi, F., Edwards, J. R., Ju, J., Bestor, T. H., and Zhang, M. Q. (2006), "Computational prediction of methylation status in human genomic sequences." *Proceedings of the National Academy of Sciences of the United States of America*, 103, 10713–6.

Davydov, E. V., Goode, D. L., Sirota, M., Cooper, G. M., Sidow, A., and Batzoglou, S. (2010), "Identifying a high fraction of the human genome to be under selective constraint using GERP++." *PLoS computational biology*, 6, e1001025.

Deaton, A. M. and Bird, A. (2011), "CpG islands and the regulation of transcription." *Genes & development*, 25, 1010–22.

Díaz-Uriarte, R. and Alvarez de Andrés, S. (2006), "Gene selection and classification of microarray data using random forest." *BMC bioinformatics*, 7, 3.

Dickson, J., Gowher, H., Strogantsev, R., Gaszner, M., Hair, A., Felsenfeld, G., and West, A. G. (2010), "VEZF1 elements mediate protection from DNA methylation." *PLoS genetics*, 6, e1000804.

Doi, A., Park, I.-H., Wen, B., Murakami, P., Aryee, M. J., Irizarry, R., Herb, B., Ladd-Acosta, C., Rho, J., Loewer, S., Miller, J., Schlaeger, T., Daley, G. Q., and Feinberg, A. P. (2009), "Differential methylation of tissue- and cancer-specific CpG island shores distinguishes human induced pluripotent stem cells, embryonic stem cells and fibroblasts." *Nature genetics*, 41, 1350–3.

Eckhardt, F., Lewin, J., Cortese, R., Rakyan, V. K., Attwood, J., Burger, M., Burton, J., Cox, T. V., Davies, R., Down, T. a., Haefliger, C., Horton, R., Howe, K., Jackson, D. K., Kunde, J., Koenig, C., Liddle, J., Niblett, D., Otto, T., Pettett, R., Seemann, S., Thompson, C., West, T., Rogers, J., Olek, A., Berlin, K., and Beck, S. (2006), "DNA methylation profiling of human chromosomes 6, 20 and 22." *Nature genetics*, 38, 1378–85.

Fan, S., Zhang, M. Q., and Zhang, X. (2008), "Histone methylation marks play important roles in predicting the methylation status of CpG islands." *Biochemical and Biophysical Research Communications*, 374, 559–564.

Fang, F., Fan, S., Zhang, X., and Zhang, M. Q. (2006), "Predicting methylation status of CpG islands in the human brain." *Bioinformatics (Oxford, England)*, 22, 2204–9.

Fogarty, J., Baker, R. S., and Hudson, S. E. (2005), "Case studies in the use of ROC curve analysis for sensor-based estimates in human computer interaction," in *GI 05 Proceedings of Graphics Interface 2005*, eds. K. Inkpen and M. Van De Panne, ACM International Conference Proceeding Series, pp. 129–136, Canadian Human-Computer Communications Society, Canadian Human-Computer Communications Society.

Gabriel, K. R. and Odoroff, C. L. (1990), "Biplots in biomedical research." *Statistics in Medicine*, 9, 469–485.

Gebhard, C., Benner, C., Ehrich, M., Schwarzfischer, L., Schilling, E., Klug, M., Dietmaier, W., Thiede, C., Holler, E., Andreesen, R., and Rehli, M. (2010), "General transcription factor binding at CpG islands in normal cells correlates with resistance to de novo DNA methylation in cancer cells." *Cancer research*, 70, 1398–407.

Good, P. J., Guyer, M. S., Kamholz, S., Liefer, L., Wetterstrand, K., Kampa, D., Sekinger, E. A., Cheng, J., Hirsch, H., Ghosh, S., Zhu, Z., Patel, S., Yang, A., Tammana, H., Bekiranov, S., Harrison, R., Church, G., Kim, T. H., Qu, C., Calcar, S. V., Luna, R., Glass, C. K., Antonarakis, S. E., Birney, E., Brent, M., Pachter, L., Reymond, A., Dermitzakis, E. T., Dewey, C., Keefe, D., Lagarde, J., Ashurst, J., Hubbard, T., Castelo, R., Sidow, A., Batzoglou, S., Trinklein, N. D., Aldred, S. F., Anton, E., Schroeder, D. I., Nguyen, L., Schmutz, J., Grimwood, J., Dickson, M., Cooper, G. M., Stone, E. A., Asimenos, G., Karnani, N., Taylor, C. M., Kim, H. K., Stamatoyannopoulos, J. A., Sabo, P., Hawrylycz, M., Humbert, R., Yu, M., Navas, P. A., Mcarthur, M., Koch, C. M., Andrews, R. M., Clelland, G. K., Wilcox, S., Fowler, J. C., Groth, P., Dovey, O. M., Ellis, P. D., Wraight, V. L., Dhami, P., Fiegler, H., Langford, C. F., Carter, N. P., Euskirchen, G., Nagalakshmi, U., Rinn, J., Popescu, G., Bertone, P., Rozowsky, J., Emanuelsson, O., Royce, T., Gerstein, M., Lian, Z., Lian, J., Nakayama, Y., Stolc, V., Tongprasit, W., Columbia, B., Agency, C., Sciences, G., Marra, M., Shin, H., and Chang, J. (2004), "The ENCODE (ENCyclopedia Of DNA Elements) Project." *Science (New York, N.Y.)*, 306, 636–40.

Guccione, E., Martinato, F., Finocchiaro, G., Luzi, L., Tizzoni, L., Dall' Olio, V., Zardo, G., Nervi, C., Bernard, L., and Amati, B. (2006), "Myc-binding-site recognition in the human genome is determined by chromatin context." *Nature cell biology*, 8, 764–770.

Hawkins, R. D., Hon, G. C., Lee, L. K., Ngo, Q., Lister, R., Pelizzola, M., Edsall, L. E., Kuan, S., Luu, Y., Klugman, S., Antosiewicz-Bourget, J., Ye, Z., Espinoza, C., Agarwahl, S., Shen, L., Ruotti, V., Wang, W., Stewart, R., Thomson, J. a., Ecker, J. R., and Ren, B. (2010), "Distinct epigenomic landscapes of pluripotent and lineage-committed human cells." *Cell stem cell*, 6, 479–91.

He, H. H. H. and Garcia, E. (2009), "Learning from Imbalanced Data," *IEEE Transactions on Knowledge and Data Engineering*, 21.

Hodges, E., Molaro, A., Dos Santos, C. O., Thekkat, P., Song, Q., Uren, P. J., Park, J., Butler, J., Rafii, S., McCombie, W. R., Smith, A. D., and Hannon, G. J. (2011), "Directional DNA Methylation Changes and Complex Intermediate States Accompany Lineage Specificity in the Adult Hematopoietic Compartment," *Molecular Cell*, 44, 17–28.

Hogart, A., Lichtenberg, J., Ajay, S. S., Anderson, S., Intramural, N. I. H., Margulies, E. H., and Bodine, D. M. (2012), "Genome-wide DNA methylation profiles in hematopoietic stem and progenitor cells reveal overrepresentation of ETS transcription factor binding sites," *Genome Research*, 22, 1407–1418.

Hon, G., Antosiewicz-bourget, J., Malley, R. O., and Castanon, R. (2011), "Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells," *Nature*, 471, 68–73.

Hua, J., Xiong, Z., Lowey, J., Suh, E., and Dougherty, E. R. (2005), "Optimal number of features as a function of sample size for various classification rules." *Bioinformatics (Oxford, England)*, 21, 1509–15.

Irzarry (2009), "Genome-wide methylation analysis of human colon cancer reveals similar hypo- and hypermethylation at conserved tissue-specific CpG island shores," *Nature genetics*, 41, 178–186.

Jaenisch, R. and Bird, A. (2003), "Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals." *Nature genetics*, 33 Suppl, 245–54.

Jones, P. a. (2012), "Functions of DNA methylation: islands, start sites, gene bodies and beyond." *Nature reviews. Genetics*, 13, 484–92.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, a. M., and Haussler, a. D. (2002), "The Human Genome Browser at UCSC," *Genome Research*, 12, 996–1006.

Kiefer, J. C. (2007), "Epigenetics in development." *Developmental dynamics : an official publication of the American Association of Anatomists*, 236, 1144–56.

Kim, S., Li, M., Paik, H., Nephew, K., Shi, H., Kramer, R., Xu, D., and Huang, T. H. (2008), "Predicting DNA methylation susceptibility using CpG flanking sequences." *Pacific Symposium On Biocomputing*, 326, 315–326.

Kim, W.-J., Kim, E.-J., Jeong, P., Quan, C., Kim, J., Li, Q.-L., Yang, J.-O., Ito, Y., and Bae, S.-C. (2005), "RUNX3 inactivation by point mutations and aberrant DNA methylation in bladder tumors." *Cancer Research*, 65, 9347–9354.

Laird, P. W. (2010), "Principles and challenges of genomewide DNA methylation analysis." *Nature reviews. Genetics*, 11, 191–203.

Larsen, F., Gundersen, G., Lopez, R., and Prydz, H. (1992), "CpG islands as gene markers in the human genome." *Genomics*, 13, 1095–107.

Lau, Q. C., Raja, E., Salto-Tellez, M., Liu, Q., Ito, K., Inoue, M., Putti, T. C., Loh, M., Ko, T. K., Huang, C., Bhalla, K. N., Zhu, T., Ito, Y., and Sukumar, S. (2006), "RUNX3 is frequently inactivated by dual mechanisms of protein mislocalization and promoter hypermethylation in breast cancer." *Cancer Research*, 66, 6512–6520.

Laurent, L., Wong, E., Li, G., Huynh, T., Tsirigos, A., Ong, C. T., Low, H. M., Wing, K., and Sung, K. (2010), "Dynamic changes in the human methylome during differentiation," *Genome Research*, 20, 320–331.

Law, J. a. and Jacobsen, S. E. (2010), "Establishing, maintaining and modifying DNA methylation patterns in plants and animals." *Nature reviews. Genetics*, 11, 204–20.

Lázcoz, P., Muñoz, J., Nistal, M., Pestaña, A., Encío, I. J., and Castresana, J. S. (2007), "Loss of heterozygosity and microsatellite instability on chromosome arm 10q in neuroblastoma." *Cancer Genetics and Cytogenetics*, 174, 1–8.

Li, Q. L., Ito, K., Sakakura, C., Fukamachi, H., Inoue, K. I., Chi, X. Z., Lee, K. Y., Nomura, S., Lee, C. W., Han, S. B., Kim, H. M., Kim, W. J., Yamamoto, H., Yamashita, N., Yano, T., Ikeda, T., Itohara, S., Inazawa, J., Abe, T., Hagiwara, A., Yamagishi, H., Ooe, A., Kaneda, A., Sugimura, T., Ushijima, T., Bae, S. C., and Ito, Y. (2002), "Causal relationship between the loss of RUNX3 expression and gastric cancer." *Cell*, 109, 113–124.

Liaw, A. and Wiener, M. (2002), "Classification and Regression by randomForest," *R News*, 2, 18–22.

Lienert, F., Wirbelauer, C., Som, I., Dean, A., Mohn, F., and Schübeler, D. (2011), "Identification of genetic elements that autonomously determine DNA methylation states." *Nature genetics*, 43, 1091–7.

Lister, R., Pelizzola, M., Dowen, R. H., Hawkins, R. D., Hon, G., Tonti-Filippini, J., Nery, J. R., Lee, L., Ye, Z., Ngo, Q.-M., Edsall, L., Antosiewicz-Bourget, J., Stewart, R., Ruotti, V., Millar, a. H., Thomson, J. a., Ren, B., and Ecker, J. R. (2009), "Human DNA methylomes at base resolution show widespread epigenomic differences." *Nature*, 462, 315–22.

Lu, L. (2010), "Predicting DNA methylation status using word composition," *Journal of Biomedical Science and Engineering*, 03, 672–676.

Macleod, D., Charlton, J., Mullins, J., and Bird, a. P. (1994), "Sp1 sites in the mouse aprt gene promoter are required to prevent methylation of the CpG island." *Genes & Development*, 8, 2282–2292.

Maunakea, A. K., Nagarajan, R. P., Bilenky, M., Ballinger, T. J., DSouza, C., Fouse, S. D., Johnson, B. E., Hong, C., Nielsen, C., Zhao, Y., Turecki, G., Delaney, A., Varhol, R., Thiessen, N., Shchors, K., Heine, V. M., Rowitch, D. H., Xing, X., Fiore, C., Schillebeeckx, M., Jones, S. J. M., Haussler, D., Marra, M. A., Hirst, M., Wang, T., and Costello, J. F. (2010), "Conserved role of intragenic DNA methylation in regulating alternative promoters," *Nature*, 466, 253–257.

Meissner, A., Mikkelsen, T. S., Gu, H., Wernig, M., Hanna, J., Sivachenko, A., Zhang, X., Bernstein, B. E., Nusbaum, C., Jaffe, D. B., Gnirke, A., Jaenisch, R., and Lander, E. S. (2008), "Genome-scale DNA methylation maps of pluripotent and differentiated cells." *Nature*, 454, 766–70.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., and Leisch, F. (2012), *e1071: Misc Functions of the Department of Statistics (e1071), TU Wien*.

Meyer, L. R., Zweig, A. S., Hinrichs, A. S., Karolchik, D., Kuhn, R. M., Wong, M., Sloan, C. A., Rosenbloom, K. R., Roe, G., Rhead, B., Raney, B. J., Pohl, A., Malladi, V. S., Li, C. H., Lee, B. T., Learned, K., Kirkup, V., Hsu, F., Heitner, S., Harte, R. A., Haeussler, M., Guruvadoo, L., Goldman, M., Giardine, B. M., Fujita, P. A., Dreszer, T. R., Diekhans, M., Cline, M. S., Clawson, H., Barber, G. P., Haussler, D., and Kent, W. J. (2013), "The UCSC Genome Browser database: extensions and updates 2013." *Nucleic acids research*, 41, D64–9.

Moayyeri, A., Hammond, C. J., Valdes, A. M., and Spector, T. D. (2012), "Cohort Profile: TwinsUK and Healthy Ageing Twin Study." *International Journal of Epidemiology*.

Previti, C., Harari, O., Zwir, I., and del Val, C. (2009), "Profile analysis and prediction of tissue-specific CpG island methylation classes." *BMC bioinformatics*, 10, 116.

Rechache, N. S., Wang, Y., Stevenson, H. S., Killian, J. K., Edelman, D. C., Merino, M., Zhang, L., Nilubol, N., Stratakis, C. a., Meltzer, P. S., and Kebebew, E. (2012),

"DNA methylation profiling identifies global methylation differences and markers of adrenocortical tumors." *The Journal of clinical endocrinology and metabolism*, 97, E1004–13.

Reich, D. E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P. C., Richter, D. J., Lavery, T., Kouyoumjian, R., Farhadian, S. F., Ward, R., and Lander, E. S. (2001), "Linkage disequilibrium in the human genome." *Nature*, 411, 199–204.

Rivenbark, A. G., Stolzenburg, S., Beltran, A. S., Yuan, X., Rots, M. G., Strahl, B. D., and Blancafort, P. (2012), "Epigenetic reprogramming of cancer cells via targeted DNA methylation." *Epigenetics official journal of the DNA Methylation Society*, 7.

Sato, K., Tomizawa, Y., Iijima, H., Saito, R., Ishizuka, T., Nakajima, T., and Mori, M. (2006), "Epigenetic inactivation of the RUNX3 gene in lung cancer." *Oncology Reports*, 15, 129–135.

Scarano, M. I., Strazzullo, M., Matarazzo, M. R., and D'Esposito, M. (2005), "DNA methylation 40 years later: Its role in human health and disease." *Journal of Cellular Physiology*, 204, 21–35.

Shen, L., Kondo, Y., Guo, Y., Zhang, J., Zhang, L., Ahmed, S., Shu, J., Chen, X., Waterland, R. a., and Issa, J.-P. J. (2007), "Genome-wide profiling of DNA methylation reveals a class of normally methylated CpG island promoters." *PLoS genetics*, 3, 2023–36.

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L. W., Richards, S., Weinstock, G. M., Wilson, R. K., Gibbs, R. A., Kent, W. J., Miller, W., and Haussler, D. (2005), "Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes," *Genome Research*, 15, 1034–1050.

Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. (2005), "ROCR: visualizing classifier performance in R." *Bioinformatics (Oxford, England)*, 21, 3940–1.

Song, J., Ugai, H., Kanazawa, I., Sun, K., and Yokoyama, K. K. (2001), "Independent repression of a GC-rich housekeeping gene by Sp1 and MAZ involves the same cis-elements." *The Journal of biological chemistry*, 276, 19897–904.

Song, J., Ugai, H., Nakata-Tsutsui, H., Kishikawa, S., Suzuki, E., Murata, T., and Yokoyama, K. K. (2003), "Transcriptional regulation by zinc-finger proteins Sp1 and MAZ involves interactions with the same cis-elements." *International Journal of Molecular Medicine*, 11, 547–553.

Stirzaker, C., Song, J. Z., Davidson, B., and Clark, S. J. (2004), "Transcriptional gene silencing promotes DNA hypermethylation through a sequential change in chromatin modifications in cancer cells." *Cancer research*, 64, 3871–7.

Tanaka, T. (2005), "[International HapMap project]." *Nihon rinsho. Japanese journal of clinical medicine*, 63 Suppl 1, 29–34.

Teschendorff, A. E., Menon, U., Gentry-Maharaj, A., Ramus, S. J., Gayther, S. a., Apostolidou, S., Jones, A., Lechner, M., Beck, S., Jacobs, I. J., and Widschwendter, M. (2009), "An epigenetic signature in peripheral blood predicts active ovarian cancer." *PloS one*, 4, e8274.

Tost, J. (2010), "DNA methylation: an introduction to the biology and the disease-associated changes of a promising biomarker." *Molecular biotechnology*, 44, 71–81.

Tsumagari, K., Baribault, C., Terragni, J., Varley, K. E., Gertz, J., Pradhan, S., Badoo, M., Crain, C. M., Song, L., Crawford, G. E., Myers, R. M., Lacey, M., and Ehrlich, M. (2013), "Early de novo DNA methylation and prolonged demethylation in the muscle lineage." *Epigenetics : official journal of the DNA Methylation Society*, 8, 317–332.

Valenzuela, L. and Kamakaka, R. T. (2006), "Chromatin insulators." *Annual review of genetics*, 40, 107–38.

Voight, B. F., Kudaravalli, S., Wen, X., and Pritchard, J. K. (2006), "A map of recent positive selection in the human genome." *PLoS biology*, 4, e72.

Weber, M., Hellmann, I., Stadler, M. B., Ramos, L., Pääbo, S., Rebhan, M., and Schübeler, D. (2007), "Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome." *Nature genetics*, 39, 457–66.

Weisenberger, D., D Siegmund, K., Campan, M., Young, J., Long, T., Faasse, M., Kang, G., Widschwendter, M., Weener, D., Buchanan, D., Koh, H., Simms, L., Barker, M., Leggett, B., Levine, J., Kim, M., French, A., Thibodeau, S., Jass, J., Haile, R., and Laird, P. (2006), "CpG island methylator phenotype underlies sporadic microsatellite instability and is tightly associated with BRAF mutation in colorectal cancer." *Nature Genetics*, 38, 787–793.

Wolffe, a. P. (1999), "Epigenetics: Regulation Through Repression," *Science*, 286, 481–486.

Zheng, H., Wu, H., Li, J., and Jiang, S.-W. (2013), "CpGIMethPred: computational model for predicting methylation status of CpG islands in human genome." *BMC medical genomics*, 6 Suppl 1, S13.

Zhou, X., Li, Z., Dai, Z., and Zou, X. (2012), "Prediction of methylation CpGs and their methylation degrees in human DNA sequences." *Computers in biology and medicine*, 42, 408–13.

Ziller, M. J., Gu, H., Müller, F., Donaghey, J., Tsai, L. T.-Y., Kohlbacher, O., De Jager, P. L., Rosen, E. D., Bennett, D. a., Bernstein, B. E., Gnirke, A., and Meissner, A. (2013), "Charting a dynamic DNA methylation landscape of the human genome," *Nature*, pp. 1–5.